Principe d'un test statistique

Pour comparer deux paramètres (deux moyennes par exemple) on va se ramener à une valeur qui suit une loi de distribution connue

Ex : N (0,1) pour la différence de deux moyennes

On va ensuite regarder sur une table de cette loi de distribution, si la valeur observée est une valeur « étonnante » ie peu probable pour cette loi de distribution « banale »

Objectif d'un test statistique

Un test permet de porter une affirmation en contrôlant le risque d'erreur

Il donne une réponse à la question :

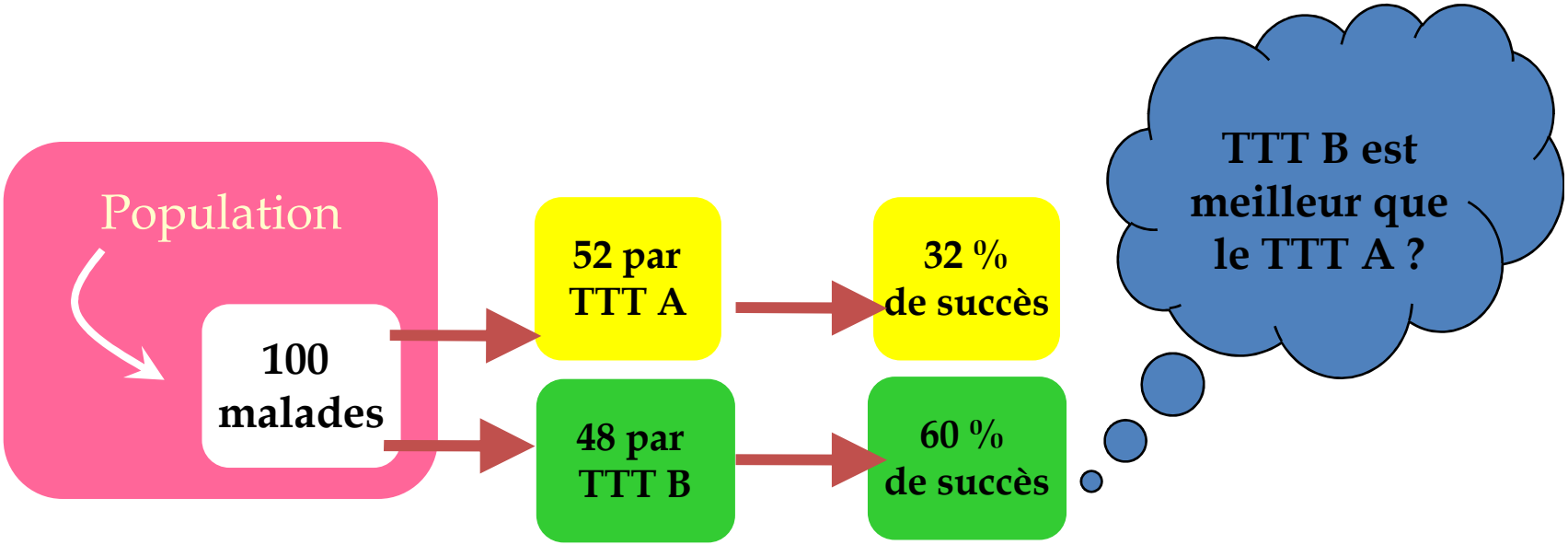
La différence observée entre mes deux paramètre peut-elle être due aux fluctuations d'échantillonnage ?

Les deux paramètres sont-ils deux estimations d'une

même population théorique ?

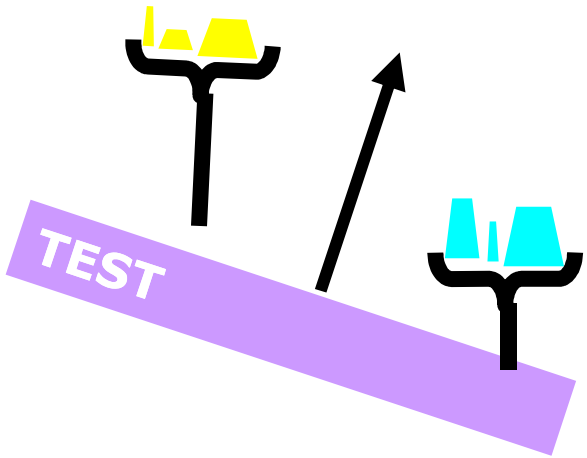
- Au moment de son examen physique annuel, Monsieur Salah présentait PA diastolique (PAD) de 97 mm Hg.
- Vous lui avez conseillé de diminuer sa consommation de sel, de perdre du poids et de faire de l'exercice.
- Bien qu'il ait fait quelques efforts dans cette direction, la PAD mesurée lors de ses trois dernières consultations était de 92, 96 et 93 mm Hg.
- Vous avez démarré un traitement médicamenteux et vous avez continué à suivre M Salah. Les mesures suivantes des PAD étaient de 88, 91 et 86 mm Hg.
- Le traitement a-t-il été efficace ?

- Quel est le meilleur traitement, A ou B ?



Analyse Univariée → **ESTIMATION**

Analyse Bivariée → **COMPARAISON**

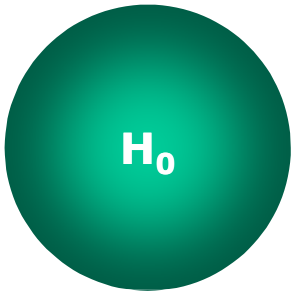


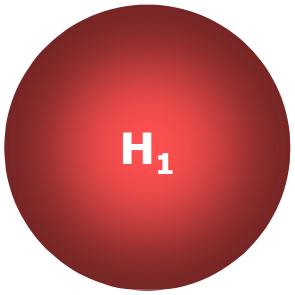
- Une comparaison porte sur des séries de données (moyennes, pourcentages, etc.)
- Test statistique = pesée
- Comparaison trouve une différence parfois grande.
- Voir si cette différence est simplement liée au hasard, ou elle est bien réelle
- Extrapoler aux populations avec un risque d'erreur

- Cette différence pourrait être expliquée par le hasard = fluctuations d'échantillonnage
- Le test statistique va permettre de quantifier le rôle du hasard dans l'observation de cette différence
- La décision est basée sur le test statistique
- La formulation de l'hypothèse nulle est la première étape du test d'hypothèse (test statistique)
- Elle peut être vérifiée
  - Rejetée
  - Gardée

Hypothèses, oui mais lesquelles ?

Nulle





Alternative

Pas de différence
Indépendance
Les paramètres d'où sont issus
les échantillons étudiés sont identiques

A = B

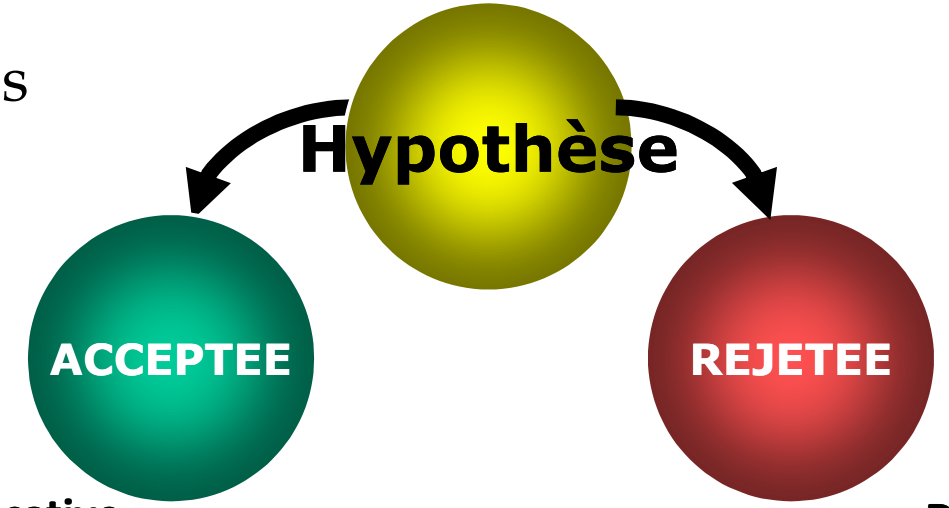
Il existe une différence
Les paramètres d'où sont
Issus les échantillons sont différents

A ≠ B
A > B
A < B

Conditions d'utilisation d'un test

- Un test n'a de sens que s'il teste une hypothèse préalablement posée

- 2 possibilités



Différence non significative

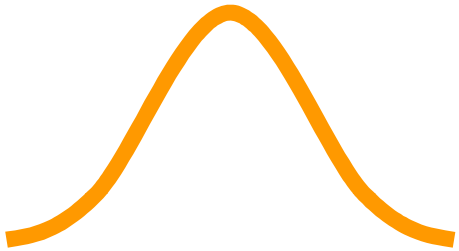
Différence significative

Contre exemple….

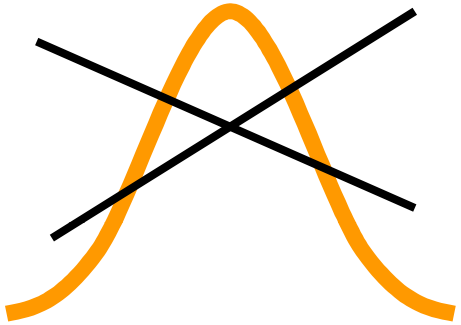
- On décide de comparer le poids des individus qui passent dimanche matin sur les trottoirs de droite et de gauche de l'avenue Didouche Mourad á Alger
- Il n'est pas impossible de trouver une différence et même parfois significative.
- Mais ceci n'aurait aucun sens et la recherche d'une explication a posteriori serait absurde

2 familles de tests.....

PARAMETRIQUE

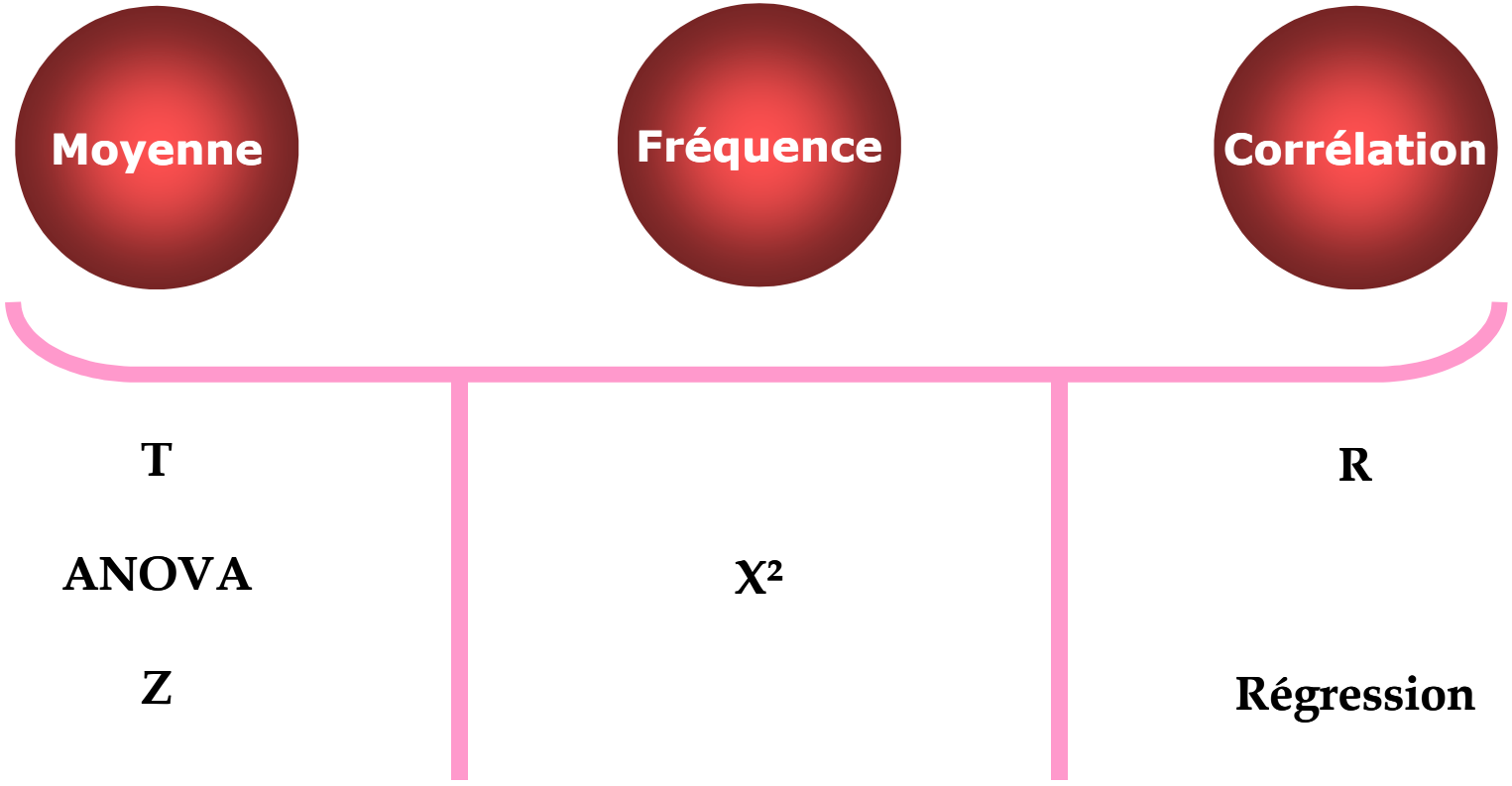


NON PARAMETRIQUE

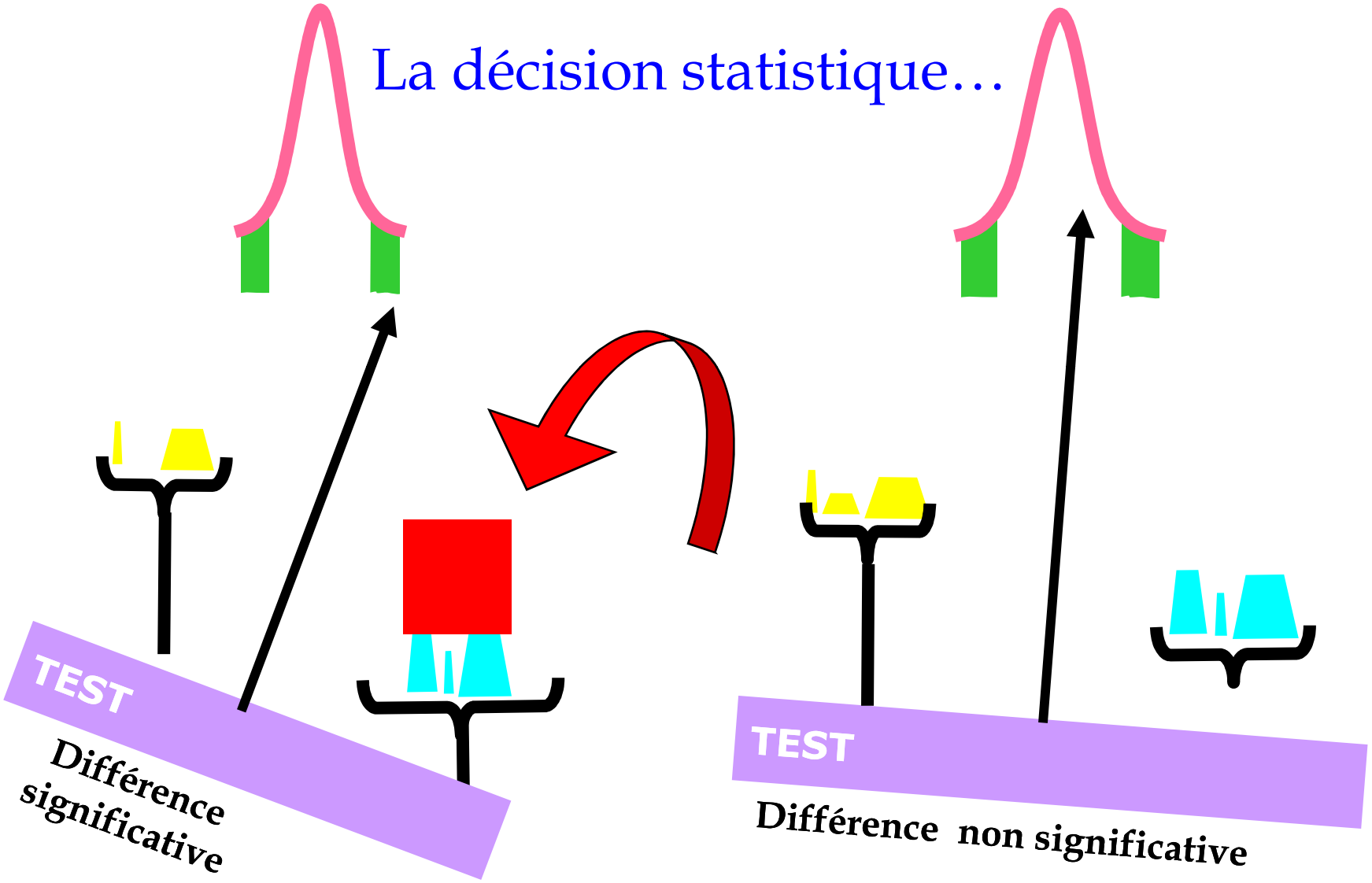


Test paramétrique	Test non paramétrique
Test t de Student	Test de Mann et Whitney
Test du Chi deux..	Test de Wilcoxon
Analyse de variance	Test de Kruskall et Wallis*
Corrélation linéaire	Test de Spearman

Tests paramétriques



La décision statistique…



Seuil de signification « p » (1)

- Le test statistique donne la probabilité « p » que le hasard puisse expliquer les résultats
- Si la probabilité « p » est inférieure ou égale à un certain seuil, appelé seuil de signification,

on rejette H0 et on dit que la différence est significative

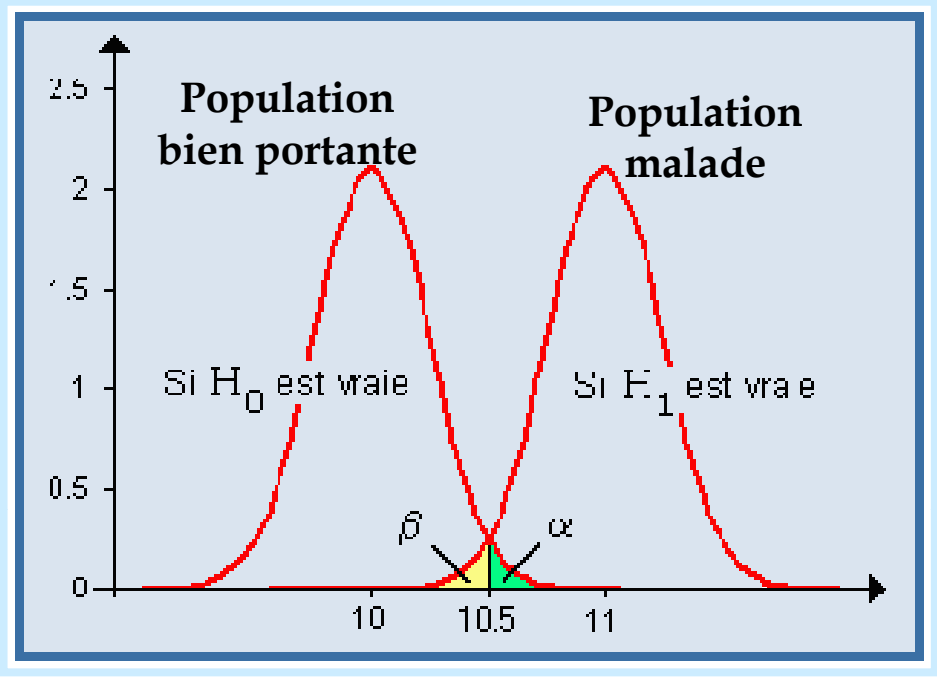
- Si « p » est supérieure au seuil,

on ne rejette pas H0 et on dit que la différence n'est pas significative

- Le seuil de signification est généralement fixé à 5 %
- Convention très largement adaptée

Seuil de signification « p » (2)

- On peut choisir un « p » plus faible que 0,05 (0.01 / 0.001…) pour réduire encore plus le rôle du hasard
- Le risque est de ne pas rejeter H0 et finalement à ne pas conclure puisqu'on aura tendance à retenir constamment H0, à moins que la probabilité « p » donnée par le test statistique ne soit elle-même pas très petite
- Le rôle du hasard ne peut être éliminé totalement.
- Le jugement est fondé sur une probabilité et n'offre aucune sécurité absolue



α = personnes bien portantes classées malades

β = personnes malades classées bien portantes

Population BORDER LINE

	H_0 vraie	H_0 fausse
H_0 acceptée	ACCORD	Erreur 2 β
H_0 rejetée	Erreur 1 α	ACCORD

Dans la prise de décision, il y a deux types d'erreurs :

α = erreur de 1è espèce

β = erreur de 2è espèce

Risque α de première espèce

- La probabilité de rejeter H0 (accepter H1), alors que H0 est vraie
- Le risque d'erreur maximal que l'on accepte en concluant à une différence qui n'existe pas.
- La valeur seuil de **α** communément admise pour rejeter H0: 5%.
- Mettre sur le marché un traitement qui n'a aucune efficacité.
- C'est un risque d'erreur grave qu'il faut savoir contrôler
- Si H0 est rejetée, il NE FAUT PAS CONCLURE : « on rejette H0 mais H0 a 5 chances sur 100 d'être vraie ».
- En effet, rejeter H0 avec un risque alpha de 5 % signifie que si H0 était vraie, la probabilité d'extraire un échantillon au moins aussi aberrant que l'échantillon observé est P = 0,05

Seuil (**α**) et degré de signification (p)

- Lorsque p ≤ **α**,

  on rejette H0 et l'hypothèse H1 est acceptée.

- Différence entre p et **α**
  - **α** est la limite sup de p que l'on accepte pour rejeter H0
  - **α** est défini *a priori* et p est calculé *a posteriori*

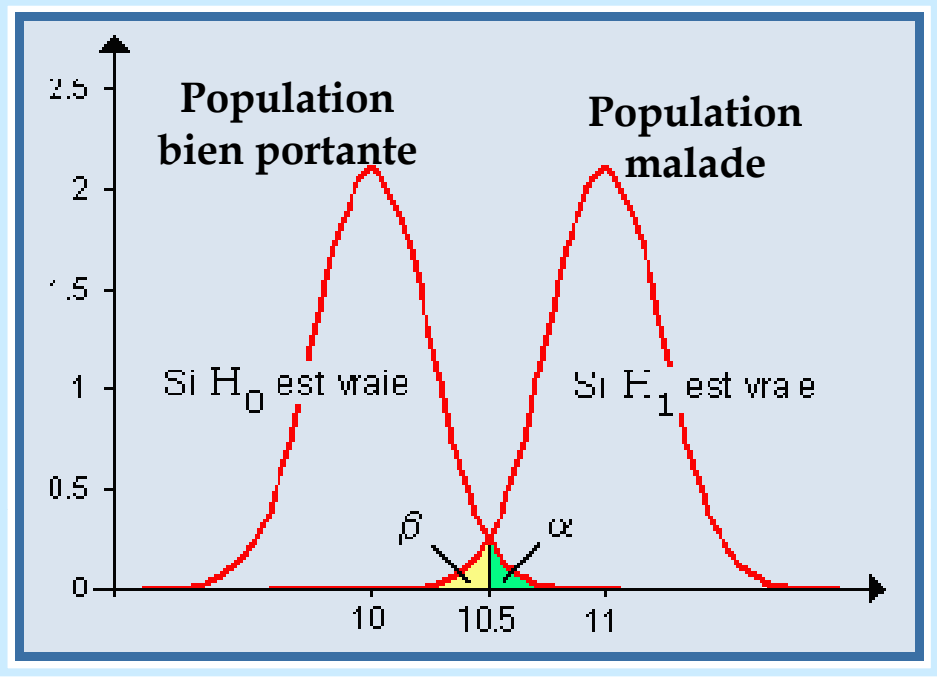
Risque β de deuxième espèce

- La probabilité d'accepter H0, alors que H1 est vraie.
- Le risque de passer à coté d'un bon traitement
- H1 = l'effet du TTT « A » diffère de celui du placebo,

ce qui correspond à une infinité de possibilités. Il n'est donc pas possible de déterminer **β** dans l'absolu.

- Le risque **β** ne peut pas être évalué lors de l'exécution du test et est exprimé par son complément à 100

La valeur 1 – **β =** puissance du test : capacité du test à détecter une différence significative lorsque celle-ci existe dans la population



α = personnes bien portantes classées malades

β = personnes malades classées bien portantes

Population BORDER LINE

	H_0 vraie	H_0 fausse
H_0 acceptée	ACCORD	Erreur 2 β
H_0 rejetée	Erreur 1 α	ACCORD

Dans la prise de décision, il y a deux types d'erreurs :

α = erreur de 1è espèce

β = erreur de 2è espèce

Antagonisme entre les deux risques

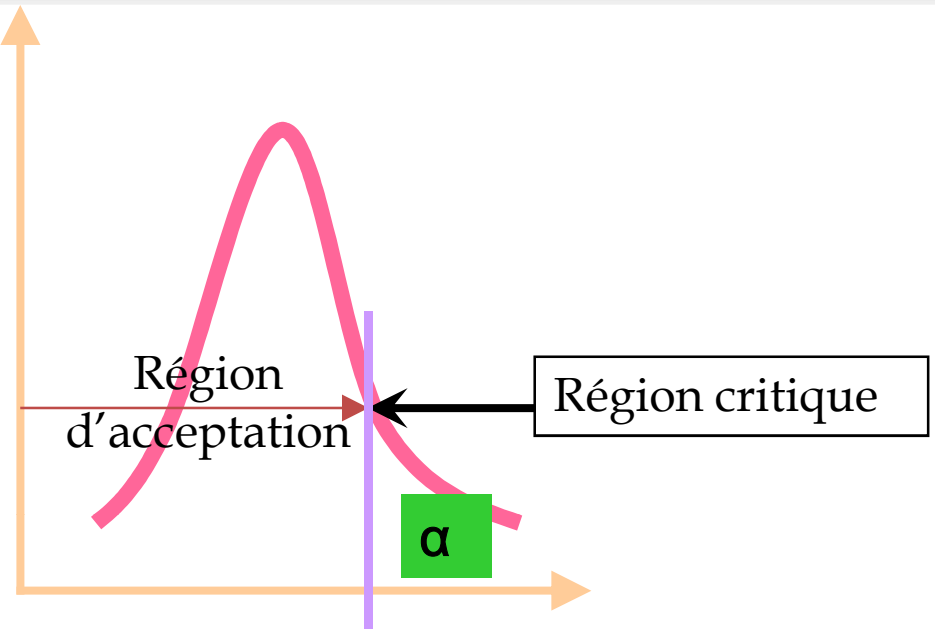
- Le risque de deuxième espèce est d'autant plus grand que le seuil de signification a été choisi petit

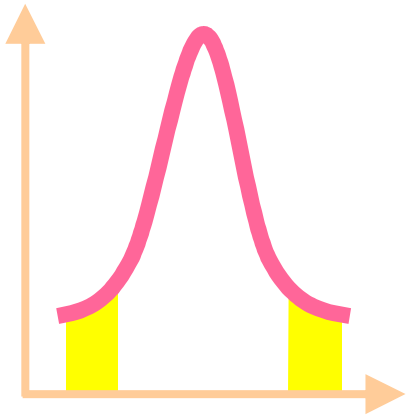
- Si on veut avoir moins de risques d'adopter un produit inactif, on a davantage de risques de laisser passer un produit actif

- Lorsqu'on rejette ou on garde une H0, on prend le risque de commettre un type d'erreur

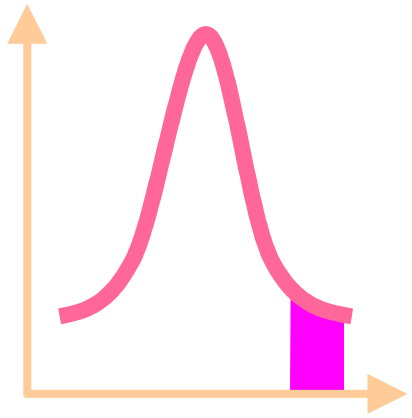
Test uni ou bilatéral ?

- H0 : Égalité des effets des traitements
- H1 = Inégalité des effets.
- Si on ne préjuge pas du sens de cette différence : le test est effectué en situation bilatérale.
- Si on croit à la supériorité d'un traitement par rapport à un autre, il faudra raisonner en terme de test unilatéral.





bilatéral



unilatéral

Application (1)

- Vous participez à la mise au point d'un nouveau traitement supposé efficace sur une maladie mortelle, mais dangereux en cas d'utilisation erronée.
- L'efficacité du traitement est testée sur des groupes de personnes saines et malades. Vous choisissez un risque alpha de
  - 10 %
  - 5 %
  - 1 %

Application (2)

- Vous participez à la mise au point d'un nouveau vaccin dans le cadre de la prévention d'une maladie grave. Ce vaccin est anodin en ce qui concerne les effets secondaires. Vous testez ce vaccin sur un échantillon versus un autre échantillon vacciné par un placebo. Vous choisissez prioritairement de diminuer
  - Le risque alpha
  - Le risque bêta
  - La puissance
  - La taille des échantillons

Au total : les étapes d'un test statistique

Nature des variables (VQN, VQL), distribution normale ?

Choix du test statistique

Définir Hypothèse nulle et alternative (H0 et H1)

Fixer le seuil de signification α et se rappeler du caractère antagoniste de β

Mécanique du calcul (indiquer le test et le calculer)

Rejeter H_0 ou pas! Et décision…

Conclusion

- Le test statistique ne peut dire que dans quelle mesure la différence observée a de chance d'être due au hasard
- Il ne juge jamais de la valeur pratique d'un traitement ou de la responsabilité étiologique d'un facteur de risque. C'est au chercheur de le faire
- Un bon protocole de recherche est nécessaire pour tirer des conclusions valides

Test de normalité

Test de Shapiro et Wilk

Ce test vérifie, si une série se distribue de façon normale

Démarche de vérification :

1/ Classer les différentes valeurs de la série par ordre croissant

Exemple : soit le tableau suivant correspondant aux résultats de mesures d'un alésage (en mm)

Pièce n°	1	2	3	4	5	6	7	8	9	10
mesure	12.124	12.230	12.327	12.242	12.466	12.215	12.026	12.359	12.215	12.387

1. Classement des valeurs de mesure par ordre croissant :

12.026	12.124	12.215	12.215	12.230	12.242	12.327	12.359	12.387	12.466

2/ Calculer la moyenne X de la série de mesure : $\overline{x} = 12.259$

3/ Calculer la variance de la série de mesure : $\delta^2 = 0.1514$

4/ Calculer les différences respectives :

$$d_1 = x_n - x_1$$

$$d_2 = x_{n-1} - x_2$$

12.026	12.124	12.215	12.215	12.230	12.242	12.327	12.359	12.387	12.466

Dans notre exemple :

$$d_1 = 12.466 - 12.026 = 0.440$$

$$d_2 = 12.387 - 12.124 = 0.263$$

$$d_3 = 12.359 - 12.215 = 0.144$$

$$d_4 = 12.327 - 12.215 = 0.112$$

$$d_5 = 12.242 - 12.230 = 0.012$$

5/ A chacune de ces différences, on affecte les coefficients a, donnés par la table, avec n nombre de différences

Dans notre exemple :

$$d_1 = 0.440 * 0.5739 = 0.2525$$

$$d_2 = 0.263 * 0.3291 = 0.0865$$

$$d_3 = 0.144 * 0.2141 = 0.0308$$

$$d_4 = 0.112 * 0.1224 = 0.0137$$

$$d_5 = 0.012 * 0.0399 = 0.0005$$

6 / Calculer la valeur :

Dans notre exemple :

$$b = \sum a_i d_j = 0.2525 + 0.0865 + 0.0308 + 0.0137 + 0.0005 = 0.384$$

7 / Calculer le rapport :

Dans notre exemple :

$$w = \frac{b^2}{\delta^2} = \frac{(0.384)^2}{0.1514} = 0.9739$$

8 / Comparer W calculé au $_{Wcritique}$ de la table, avec n nombre de données.

Si W calculé est supérieur au $_{Wcritique}$ de la table, la normalité est acceptée.

Si W calculé est inférieur au $_{Wcritique}$ de la table, la normalité est rejetée

Dans le cas de l'exemple, W = 0.9739 > 0.842

l'hypothèse de normalité est acceptée.

(Si W < 0.842, il y aurait refus avec un risque de 5% de rejeter une distribution normale.)

n	2	3	4	5	6	7	8	9	10	
J										
1	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739	
2		0.0000	0.1677	0.2413	0.2806	0.3031	0.3164	0.3244	0.3291	
3				0.0000	0.0875	0.1401	0.1743	0.1976	0.2141	
4						0.0000	0.0561	0.0947	0.1224	
5								0.0000	0.0399	

n	11	12	13	14	15	16	17	18	19	20
J										
1	0.5601	0.5475	0.5359	0.5251	0.5150	0.5056	0.4963	0.4886	0.4808	0.4734
2	0.3315	0.3325	0.3325	0.3318	0.3306	0.3290	0.3273	0.3253	0.3232	0.3211
3	0.2260	0.2347	0.2412	0.2460	0.2495	0.2521	0.2540	0.2553	0.2561	0.2565
4	0.1429	0.1586	0.1707	0.1802	0.1878	0.1939	0.1988	0.2027	0.2059	0.2085
5	0.0695	0.0922	0.1099	0.1240	0.1353	0.1447	0.1524	0.1587	0.1641	0.1686
6	0.0000	0.0303	0.0539	0.0727	0.0880	0.1005	0.1109	0.1197	0.1271	0.1334
7			0.0000	0.0240	0.0433	0.0593	0.0725	0.0837	0.0932	0.1013
8					0.0000	0.0196	0.0359	0.0496	0.0612	0.0711
9							0.0000	0.0163	0.0303	0.0422
10									0.0000	0.0140

N	W '95%'	W '99%'
10	0.842	0.781
11	0.850	0.792
12	0.859	0.805
13	0.856	0.814
14	0.874	0.825
15	0.881	0.835
16	0.837	0.844
17	0.892	0.851
18	0.897	0.858
19	0.901	0.863
20	0.905	0.868
21	0.908	0.873
22	0.911	0.878
23	0.914	0.881
24	0.916	0.884
25	0.918	0.888
26	0.920	0.891
27	0.923	0.894
28	0.924	0.896
29	0.926	0.898
30	0.927	0.900
31	0.929	0.902
32	0.930	0.904

TESTS PARAMETRIQUES

Le terme communément utilisé « tests paramétriques » recouvre les tests statistiques fondés sur des hypothèses sur la loi de distribution (répartition) de la variable étudiée.

Il existe de nombreuses lois de distribution que l'on peut résumer par certaines valeurs caractéristiques encore appelées paramètres, d'où ce terme de « tests paramétriques ».

Dans la majorité des cas, ces tests paramétriques sont basés sur la loi normale, qui possède deux paramètres : la moyenne et l'écart-type qui suffisent à connaître la loi de probabilité de distribution.

les conditions d'application du test : normalité et égalité des variances. On a déjà vu qu'il était très difficile, surtout avec de si petits échantillons, de vérifier la normalité.

TEST DE STUDENT

Ce test permet de comparer deux distributions extraites d'une population normale ou approximativement normale au niveau de leurs moyennes.

Il s'agit de décider si la différence observée entre les moyennes des deux échantillons de comparaison est attribuable à la variable indépendante testée ou si elle peut être considérée comme l'effet du hasard.

1/ Cas de deux échantillons indépendants

Deux séries de mesure pour lesquelles il n'y a aucune correspondance entre les éléments de la première série et ceux de la deuxième; les deux séries de mesures sont obtenues avec des sujets différents. Dans ce cas le but de l'application du test t est de voir si les deux moyennes calculées sur les deux échantillons diffèrent significativement.

Soit la situation suivante :

HYPOTHESE

H0: m1 = m2 (c'est-à-dire les deux groupes de comparaison appartiennent à des populations qui possèdent des moyennes identiques)

H1: m1 ≠ m2 () ou m1 < m2 ou m1 > m2 (Hypothèses unilatérales)

hypothèse bilatérale

Conditions d'application

La distribution des données de chaque échantillon ne peut pas différer fortement de la normale, et, en particulier, ne pas être trop dissymétrique, surtout si les échantillons sont petits

- Les variances des populations de provenance ne peuvent pas être extrêmement différentes

- Les tailles des échantillons ne peuvent pas extrêmement différentes

$$t = \frac{m_1 - m_2}{\sqrt{V_C\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

m_1: moyenne du premier échantillon

m_2: moyenne du deuxième échantillon

n_1: nombre de mesures (sujets) du premier échantillon

n_2: nombre de mesures (sujets) du deuxième échantillon

Vc: la variance commune, c'est une sorte de moyenne des deux variances (V1 et V2) pondérée par le nombre de mesures n1 et n2; sa formule est:

$$n_1 = n_2 \Rightarrow V_C = \frac{V_1 + V_2}{2}$$

$$n_1 \neq n_2 \Rightarrow V_C = \frac{V_1(n_1 - 1) + V_2(n_2 - 1)}{n_1 + n_2 - 2}$$

Une fois on a la valeur de t calculé, se rapporter à la table de t de Student pour comparer le "t calculé" au "t critique" , et ce,

au degré de liberté = [(n1+ n2) - 2] et au seuil $\alpha = 0{,}05$.

La différence est significative si

"t calculé" est supérieur ou égal au "t critique".

Exemple

Soit deux groupes de sujets ayant subi une expérience sur la mémoire (retenir une série de mots).
n1 =27 ont obtenu une moyenne de 63,5, une médiane = 63 et un écart-type de 15, 6
n2= 18 ont obtenu une moyenne de 48,7, une médiane= 49 avec un écart-type de 16,4
Question: y a-t-il une différence entre la performance de deux groupes?

Pour pouvoir appliquer t de Student on doit vérifier la normalité de deux distribution, l'homogénéité des variances et calculer la variance commune.

$$CD_1 = \frac{3(\mu_1 - M_d)}{\partial_1} = \frac{3(63.5 - 63)}{15.6} = 0.096$$

$$CD_2 = \frac{3(\mu_2 - M_d)}{\partial_2} = \frac{3(48.7 - 49)}{16.4} = -0.054$$

Vérification de la normalité des distributions

$$CD_1 = \frac{3(\mu_1 - M_d)}{\partial_1} = \frac{3(63.5 - 63)}{15.6} = 0.096$$

$$CD_2 = \frac{3(\mu_2 - M_d)}{\partial_2} = \frac{3(48.7 - 49)}{16.4} = -0.054$$

Vérification de l'homogénéité des variances

$$F_{Cal} = \frac{(16.4)^2}{(15.6)^2} = 1.105$$

La valeur critique de F à ddl horizontal = 17 (18-1) et ddl vertical = 26 (27-1) égal 1,89 (17 et 26 ne figurent pas sur la table, on prendra les valeurs immédiatement supérieures.

F calculé étant inférieur à F critique, on conclue donc que la différence entre les variance n'est pas significative

$$V_C = \frac{(15.6)^2(27-1)+(16.4)^2(18-1)}{27+18-2} = 253.48$$

$$t = \frac{63.5 - 48.7}{\sqrt{253.48\left(\frac{1}{27}+\frac{1}{18}\right)}} = 3.06$$

Nous devons maintenant chercher la valeur critique de t.

ddl = 27+18-2 = 43; au seuil 0,05 t = 2,02 (43 n'existe pas sur la table, on choisira le degré de liberté juste inférieur c'est-à-dire 40).

3,06 étant > 2,02,

nous rejetons H0 et nous admettons l'existence d'une différence significative entre m1 et m2.

Cas de deux échantillons dépendants ou appareillés

Il s'agit de deux séries de mesures pour lesquelles il y a une correspondance stricte, terme à terme, entre les éléments de l'une et les éléments de l'autre. C'est le cas par exemple de deux séries de notes. relevées auprès d'un échantillon d'élèves, la première avant les vacances et la deuxième à la rentrée; il y a donc une correspondance parfaite puisque c'est le même groupe qui effectue les deux épreuves.

Là encore on va calculer un t qui indique si les deux moyennes sont significativement différentes. La formule sera légèrement modifiée par rapport à la précédente:

$$t = \frac{m_d}{\partial_d / \sqrt{N}} = \frac{\sum |m_1 - m_2|}{\partial_d / \sqrt{N}}$$

md: la moyenne des différences
σd: l'écart-type de la distribution des différences
N: le nombre de sujets

Exemple

dans une expérience sur la perception du langage, on fait subir deux épreuves à un même groupe de 40 sujets. On a obtenu les mesures suivantes qui désignent le nombre de mots correctement reproduits:

Sujets	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Epreuve1	3	5	5	7	7	7	4	6	6	7	4	8	5	8	6	8	6	7	6	7
Epreuve2	5	2	4	2	6	3	4	1	3	4	1	3	3	2	5	3	2	7	3	3

Sujets	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**
Epreuve1	9	7	6	5	7	5	7	9	6	7	6	6	8	4	6	4	8	5	5	8
Epreuve2	3	3	5	2	4	2	6	3	2	4	3	5	2	4	2	4	5	3	4	4

On commence par calculer les différences entre les mesures de l'épreuve1 et celles de l'épreuve2
puis les relever au carré (d^2). On obtient la distribution suivante:

Sujets	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
d	-2	+3	+1	+5	+1	-4	0	+5	+3	+3	+3	+5	+2	+6	+1	+5	+4	0	+3	+4
d²	4	9	1	25	1	46	0	25	9	9	9	25	4	30	1	25	16	0	9	16

Sujets	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**
d	+6	+4	+1	+3	+3	+3	+1	+6	+4	+3	+3	+1	+6	0	+4	0	+3	+2	+1	+4
d²	36	16	1	9	9	9	1	36	16	9	9	1	36	0	16	0	9	4	1	16

2- on calcule ensuite la somme des différences (∑d) et la moyenne de ces différences (md)

$$m_d = \frac{114}{40} = 2.85$$

on calcule la variance puis l'écart-type de cette distribution des différences (Vd et σd)

$$V_d = \frac{\sum d^2 - \frac{(\sum d)^2}{N}}{N-1} = \frac{474 - \frac{(114)^2}{40}}{39} = 3.82$$

L'écart-type des différences

$$\partial_d = \sqrt{V_d} = \sqrt{3.82} = 1.95$$

4- On applique la formule de t pour échantillons appareillés

$$t = \frac{m_d}{\partial_d / \sqrt{N}} = \frac{2.85}{1.95 / \sqrt{40}} = 9.23$$

dans la table de t, on cherche la valeur critique au ddl N-1 = 40-1 =39 et au seuil 0,05; on trouve t= 2,02 ce qui est largement inférieur à t calculé,

la différence entre les deux moyennes est donc très significative

Test de Khi-deux (X^2)

Il permet de comparer deux ou plusieurs groupes caractérisés par une répartition de leurs effectifs respectifs.

1) Cas des échantillons indépendants

1. Ce test n'est applicable que si les catégories sont les mêmes dans les différents échantillons

2. Les données doivent être indépendantes d'une colonne à l'autre ou d'une rangée à l'autre (pas d'échantillons appareillés).

3. Les groupes doivent avoir une taille suffisante, ce test ne pas être appliqué si 20% ou plus des fréquences attendues sont inférieures à 5, sinon il faut apporter la correction de Yates.

Test de Khi-deux (X^2)

Calculer l'effectif théorique pour chaque case

Calculer la statistique khi-deux pour chaque case

Faire la somme des khi-deux obtenus

Comparer ce résultat avec la valeur tabulaire correspondant au seuil de signification choisi et au nombre de degré de liberté que comporte la situation. Si le résultat est supérieur ou égal à cette valeur, alors on rejette H0

Soit une variable nominale trichotomique VA formée de 2 modalités: a1 et a2

Soit une variable ordinale de catégories rangées VB à 3 modalités: b1; b2 et b3

1/Dresser le tableau des effectifs observés

	b1	**b2**	**b3**	**Total**
a1	n1	n2	n3	L1
a2	n4	n5	n6	L2
Total	C1	C2	C3	**N**

Test de Khi-deux (X^2)

Calculer les effectifs théoriques (appelés également attendus)

	b1	**b2**	**b3**	**Total**
a1	n'1	n'2	n'3	L1
a2	n'4	n'5	n'6	L2
Total	C1	C2	C3	**N**

L:Total ligne

C:Total colonne

N: Effectif total

	b1	**b2**	**b3**
a1	n'1=$C1 \times L1/N$	n'2= $C2 \times L1 / N$	n'3= $C3 \times L1 / N$
a2	n'4= $C1 \times L2/N$	n'5=$C2 \times L2/N$	n'6= $C3 \times L2/N$

Calculer le Khi-deux des cases

Pour chaque case, on applique: (effectif observé – effectif théorique)² /effectif théorique

	b1	**b2**	**b3**
a1	$(n1-n'1)^2/n'1$	$(n2-n'2)^2/n'2$	$(n3-n'3)^2/n'3$
a2	$(n4-n'4)^2/n'4$	$(n5-n'5)^2/n'5$	$(n6-n'6)^2/n'6$

Test de Khi-deux (X^2)

Si 20% au plus des effectifs théoriques sont inférieurs à 5, on apporte la correction de Yates et la formule devient: (|effectif observé – effectif théorique| - 0.5) / effectif théorique

Calculer le Khi-deux (la somme de chacune des cases de l'étape précédente).

Déterminer les degrés de liberté de Khi-deux en appliquant la formule suivante:

ddl=(Nombre de colonnes – 1) (Nombre de lignes – 1)

La valeur de Khi-deux calculée doit être comparée à la valeur critique de Khi-deux (sur la table)

au seuil 0,05: si Khi-deux calculé est supérieur au Khi-deux théorique, on considère la différence significative, c'est-à-dire qu'il y a influence de la variable indépendante sur la variable dépendante.

Remarque: cette procédure est générale, qu'il s'agisse d'un tableau à quatre cases ou plus.

Exemple

Test de Khi-deux (X^2)

Question problème: le choix de la filière dépend t-il de la catégorie socioprofessionnelle?

H0: Il n'y a pas effet de la catégorie socioprofessionnelle sur le choix de la filière

H1: La catégorie socioprofessionnelle influence le choix de la filière.

1/ Effectifs observés

	Maths	Lettres	Sciences	Technique	Total
Très défavorisée	7	2	1	3	13
Défavorisée	6	3	3	0	12
Moyenne	4	3	3	2	12
Favorisée	5	1	6	1	13
Total	22	9	13	6	**50**

2/ Effectifs théoriques

	Maths	Lettres	Sciences	Technique	Total
Très défavorisée	5.72	2.34	3.38	1.56	13
Défavorisée	5.28	2.16	3.12	1.44	12
Moyenne	5.28	2.16	3.12	1.44	12
Favorisée	5.72	2.34	3.38	1.56	13
Total	22	9	13	6	50

Test de Khi-deux (X^2)

Khi-2 des cases

Puisque 12 fréquences théoriques sur 16 sont inférieures à 5, on applique la correction de Yates on obtient le tableau suivant:

$$\chi^2_{yates} = \sum^{k} \frac{(|f_0 - f_{th}| - 0.5)^2}{f_{th}}$$

	Maths	Lettres	Sciences	Technique
Très défavorisée	0.10	0.01	1.04	0.56
Défavorisée	0.009	0.05	0.04	0.61
Moyenne	0.11	0.05	0.04	0.0025
Favorisée	0.008	0.3	1.32	0.002

$$\chi^2_{cal} = 4.25$$

Test de Khi-deux (X^2)

$$ddl = (4-1)*(4-1) = 9$$

$$\chi^2_{critique} = 16.9$$

$$\alpha = 0.05$$

$4.25 < 16.9$, donc H0 est retenue.

La catégorie socio professionnelle n'a pas d'effet sur le choix de la filière.

Test de Khi-deux (X^2)

Cas des échantillons dépendants

Il s'agit de comparer un tableau de fréquences construit sur des dichotomies (fréquences recueillies auprès d'un seul échantillon à des moments différents ou dans deux situations différentes).

Supposant, par exemple que l'on veuille étudier la différence entre le nombre d'élèves accédant à deux types de formation

FORMATION B	ADMIS	REFUSES	TOTAL
ADMIS	n1	n2	N1
REFUSES	n3	n4	N2
TOTAL	N3	N4	N

FORMATION A

Ce sont les mêmes candidats (ayant participé à l'examen de la formation A et l'examen de la formation B),

on veut comparer la proportion des admis à la première formation avec la proportion des admis à la deuxième formation)

c'est - à - dire les fréquences :

$$P_1 = \frac{N_3}{N} \text{ et } P_2 = \frac{N_1}{N}$$

Pour ce faire, on calcule un χ^2 assez différent du précédent : $\chi^2 = \frac{(n_2 - n_3)^2}{n_2 + n_3}$

Remarquons que cette formule ne s'intéresse qu'aux effectifs des cases hétérogènes (admis à une formation et refusés à une autre).

Exemple (tiré de S. Ehrlich et C. Flament, 1970, p.158):

On a posé à 300 personnes deux questions: "allez-vous souvent au cinéma?" et "allez-vous souvent au théâtre?".

Les réponses sont "oui" ou "non". On observe les résultats suivants:

		CINEMA		
		OUI	NON	TOTAL
THEATRE	OUI	n1=42	n2=48	N1=90
	NON	n3=78	n4=132	N2=210
	TOTAL	N3=120	N4=180	N=300

42 personnes vont souvent au **cinéma** et au **théâtre**;

78 personnes vont souvent au cinéma et rarement au théâtre;

120 personnes vont souvent au cinéma;

90 personnes vont souvent au théâtre

La question: la différence entre ces deux nombres est-elle significative?

on calcule un χ^2 assez différent du précédent :

$$\chi_{Cal}^{\ 2} = \frac{(n_2 - n_3)^2}{n_2 + n_3} = \frac{(48-78)^2}{45+78} = \frac{900}{126} = 7.14$$

La valeur critique $\chi_{Cri}^{\ 2}$ pour ddl = 1 et 0,05 probabilité d'erreur donne 3.84

la différence est dons significative.

Tableau I.
Principaux tests paramétriques.

Situations	Tests	Conditions
A – Liaison de 2 variables qualitatives : (comparaison de 2 pourcentages)		
1 – Indépendantes	a – Test de l'écart-réduit pour des pourcentages (non utilisé)	
	b – Test du Chi^2(*)	si effectifs théoriques ≥ 5
	c – Test du Chi^2 corrigé de Yates (*)	si effectifs théoriques ≥ 3 ou < 5 sinon test de Fisher
2 – Appariées	Test du Chi^2 de Mac Némar (*)	si paires discordantes ≥ 10 (ddl=1)
B – Liaison entre 1 variable qualitative à 2 classes et 1 variable quantitative (comparaison de 2 moyennes) :		
1 – échantillons indépendants	a – Test de l'écart-réduit	effectifs des 2 échantillons ≥ 30
	b – Test de Student	effectif d'un échantillon < 30 Loi normale égalité des variances
2 – échantillons appariés	a – Test de l'écart-réduit apparié	si effectif ≥ 30
	b – Test de Student apparié	si effectif < 30 distribution normale des différences entre les deux échantillons pour chaque individu
C – Liaison de 1 variable qualitative à n > 2 classes à 1 variable quantitative (comparaison de n (> 2) moyennes)		
1 – Indépendantes	Analyse de variance	variances égales Loi normale
2 – Appariées	Analyse de variance à deux facteurs	variances égales Loi normale
D – Liaison entre 2 variables quantitatives	Coefficient de corrélation	distribution liée normale et variance constante

(*) Le chi^2 est un test semi-paramétrique.

les tests non paramétriques

Il existe de tests moins "exigeants" en conditions d'applications, notamment en ce qui concerne la taille de l'échantillon, la normalité de la distribution et l'égalité des variances, ces tests sont dits non paramétriques.
Le principe de base de ces tests est de transformer les données en rangs et à mesurer

Parmi ces tests nous allons voir:

le test de Mann-Wihtney, l'alternative non paramétrique de t de Student pour deux échantillons indépendants;

le test de Wilcoxon, l'alternative non paramétrique de t de Student pour deux échantillons dépendants;

le test de Kruskal-Wallis, l'alternative non paramétrique de l'analyse de variance.

U de Mann-Withney

Ce test est destiné à étudier si une variable indépendante nominale dichotomique influence une variable dépendante ordinale de scores rangés ou d'intervalle.

Ce test doit être préféré au test t de student lorsque la distribution n'obéit pas à la loi normale (donc remarquablement dissymétrique)

U de Mann-Withney

Algorithme de résolution

$$\text{cas : } n_A \text{ et } n_B \text{ sont supérieurs à } 8$$

Supposant les donnés suivantes:

A	11	9	7	12	12	40	5	4	15	10	10	14
B	13	15	15	14	35	18	13	25	20	6	5	

$n_A = 12$

$n_B = 11$

1. Mélanger les donnés de deux groupes
2. Ordonner la série obtenue en ordre croissant
3. Accorder des rangs; pour les ex-æquo attribuer à chacun le rang moyen
4. Reconstruire les deux groupes avec données et les rangs correspondants

U de Mann-Withney

calculer U et U'

$$U = n_A n_B + \frac{n_A(n_A+1)}{2} - T_A$$

$$U' = n_A n_B + \frac{n_B(n_B+1)}{2} - T_B$$

Mann et Whitney ont montré que la variable U se distribue selon une loi approximativement normale.

Calculer donc: la moyenne et l'écart-type de la distribution de U:

$$m_U = \frac{n_A * n_B}{2} \qquad \delta_U = \sqrt{\frac{(n_A * n_B)(n_A + n_B + 1)}{12}}$$

Il en est de même pour U', valeur symétrique de U. il suffit donc de tester l'écart entre U et m (ou entre m et U')

$$|Z| = \frac{|U - m_U|}{\delta_U}$$

Si nous revenons à notre exemple, nous aurons donc:

U de Mann-Withney

Scores	4	5	5	6	7	9	10	10	11	12	12	13
Rangs	1	2.5	2.5	**4**	**5**	6	**7.5**	7.5	9	10.5	10.5	**12.5**
Scores	13	14	14	15	15	15	18	20	25	35	40	
Rangs	**12.5**	14.5	14.5	**17**	17	17	**19**	20	21	22	**23**	

A	Scores	11	9	7	12	12	40	5	4	15	10	10	14
	Rangs	9	6	5	**10.5**	**10.5**	23	**2.5**	1	17	7.5	7.5	**14.5**
B	Scores	13	15	15	14	35	18	13	25	20	6	5	
	Rangs	**12.5**	17	17	**14.5**	22	19	**12.5**	21	20	4	**2.5**	

calculer la somme des rangs de A et la somme des rangs de B

$T_A = 114$

$T_B = 162$

$$U = n_A n_B + \frac{n_A(n_A+1)}{2} - T_A = (12*11) + \frac{12*13}{2} - 114 = 96$$

$$\mathrm{U}' = n_A n_B + \frac{n_B(n_B+1)}{2} - T_B = (12*11) + \frac{11*10}{2} - 162 = 36$$

$$m_U = \frac{12*11}{2} = 66 \quad \delta_{\mathrm{U}} = \sqrt{\frac{(12*11)(12+11+1)}{12}} = 16.25$$

On peut vérifier que : $\frac{\mathrm{U}+\mathrm{U}'}{2} = \frac{96+66}{2} = 66 = m_U$

Par conséquent : $|\mathrm{Z}| = \frac{|96-66|}{16.25} = 1.85$

Vérifier la signification de la valeur Z :

Si Z calculé est supérieur ou égal à 1.96, la différence est significative au $\mathrm{P} = 0.05$

Si Z calculé est supérieur ou égal à 2.56, la différence est significative au $\mathrm{P} = 0.01$

Algorithme de résolution

$$\text{cas : } n_A \text{ et } n_B \text{ sont inférieurs à } 8$$

Dans ce cas, la distribution n'est pas gaussienne, le modèle précédent ne peut pas être appliqué.

Mann et Withney ont construit des tables avec des valeurs critiques qu'il est possible de consulter directement en fonction de:

de **U** si U est inférieur à U'

de **U'** si U' est inférieur à U

Supposons les mesures de deux groupes et leurs rangs:

A	Scores	6	3	10	5	14	$n_A = 5$
	Rangs	7	9	5	**8**	**3**	$T_A = 32$
B	Scores	12	8	16	18		$n_B = 4$
	Rangs	**4**	6	2	**1**		$T_B = 13$

$$U = 5*4 + \frac{5(5+1)}{2} - 32 = 3$$

$$U' = 5*4 + \frac{4(4+1)}{2} - 13 = 17$$

La table est consultée en fonction de l'effectif n2 du plus grand de deux échantillons (ici, n2 =5).

Pour n1=4 et U=3, nous lisons P=.056

L'hypothèse nulle n'est pas rejetée, il n'existe pas une différence entre les moyennes des rangs.

Remarque:

Pour simplifier les calculs on prendra toujours la somme des rangs dans la situation comportant le moins de sujets.

Lorsqu'il y aura le même nombre de sujets dans les deux conditions, il sera possible de prendre l'une ou l'autre des deux conditions pour calculer la somme des rangs.

Le test de Kruskal-Wallis

C'est la généralisation du test de Mann-Withney à trois échantillons ou plus.

Les scores sont remplacés par les rangs obtenus à l'intérieur d'un seul groupe constitués à partir des échantillons à comparer.

Supposant 4 groupes de sujets reçoivent un enseignement selon quatre méthodes différentes.

On souhaite comparer leurs résultats sur la base des données suivantes:

Groupes	1	2	3	4
Scores	8	15	18	4
	20	14	16	7
	13	7	15	12
	14	9	19	10
	17	12		8
		10		6
				11
Effectifs	$n_1=5$	$n_2=6$	$n_3=4$	$n_4=7$

On mélange les 4 groupes (k=4) et on ordonne les scores: 4 - 6 - 7 - 7 - 8 - 8 - 9 - 10 - 10 - 11 - 12 - 12 - 13 - 14 - 14 - 15 - 15 - 16 - 17 - 18 - 19 - 20

Groupe1		Groupe2		Groupe3		Groupe4	
Scores	Rangs	Scores	Rangs	Scores	Rangs	Scores	Rangs
8	5.5	15	16.5	18	20	4	1
20	22	14	14.5	16	18	7	3.5
13	13	7	3.5	15	16.5	12	11.5
14	14.5	9	7	19	21	10	8.5
17	19	12	11.5			8	5.5
		10	8.5			6	2
						11	10
$T_1=74$ $m_1=14.8$		$T_2=61,5$ $m_2=10,25$		$T_3=75.5$ $m_2=18.875$		$T_4=42$ $m_4=6$	

Notons que $\sum T_i$ (somme des totaux des rangs) $= \frac{N(N+1)}{2}$

$$74 + 61.5 + 75.5 + 42 = \frac{22(22+1)}{2} = 235$$

On applique la formule de Kruskal et Wallis:

$$H = \left[\frac{12}{N(N+1)} \times \frac{\sum T_i^2}{n_i} \right] - 3(N+1)$$

$$H = \frac{12}{22(22+1)} \times \left[\frac{(74)^2}{5} + \frac{(61.5)^2}{6} + \frac{(75.5)^2}{4} + \frac{(42)^2}{7} \right] - 3(22+1) = 11.64$$

Cette variable H suit une loi de X^2

. Il suffit donc de revenir à la table de X^2 et de comparer H calculé à la valeur critique de X^2 au ddl = k – 1 (c'est-à-dire nombre de groupes – 1).

K = 4; k – 1= 3. La valeur critique de X^2 au ddl = 3 et P=0,05 est égale à 7,82.

La valeur calculée est supérieure à la valeur théorique, on rejette donc H0.

Autrement dit il existe des différences entre

les moyennes des rangs des 4 groupes.

Le test de Wilcoxon

Il permet la comparaison les moyennes des rangs de deux échantillons appareillés.

Son principe consiste à classer les sujets dans l'ordre croissant des valeurs absolues des différences non nulles.

Supposons les données suivantes portant sur les notes (variant de 0 à 10) obtenues par un groupe d'élèves à deux moments différents de l'année scolaire:

Eléves	A	b	c	d	e	f	g	h	i	j
Première note(A)	1	1	2	2	7	2	3	6	4	5
Deuxième note(B)	9	6	9	2	5	7	8	8	7	4

calculer pour chaque élève la différence entre la première et la deuxième note

$$D=B-A$$

ce qui donnera la distribution suivante

Eléves	a	b	c	d	e	f	g	h	i	j
D	+8	+5	+7	0	-2	+5	+5	+2	+3	-1

Nous constatons que:

Un élève n'a ni régressé ni progressé (d=0)

Deux élèves ont régressé (d négative)

Sept élèves ont progressé (d positive)

Classer les sujets dans l'ordre croissant des valeurs absolues des différences non nulles c'est-à-dire dans cet exemple 9 valeurs

(on élimine l'élève **d** car sa différence =0)

Eléves	a	b	c	d	e	f	g	h	i	j
D	+8	+5	+7	0	-2	+5	+5	+2	+3	-1

D	1	2	3	5	7	8
sujets	j	e-h	i	b–f-g	c	a
rangs	1	2,5	4	6	8	9

calculer la sommes des rangs des différences positives (T^+)

calculer la sommes des rangs des différences négatives (T^-)

Dans notre exemple T+ est la somme des différences des sujets **a, b, c, f, g, h, i**.

T- est la somme des différences des sujets **e , j**

T+=41.5 T-=3.5

$$T^{+} = 41.5 \qquad T^{-} = 3.5$$

$$\text{Notons que : } T^{+} + T^{-} = \frac{n(n+1)}{2}$$

$$41.5 + 3.5 = \frac{9(9+1)}{2} = 45$$

tester la plus petite valeur des T^{+} T^{-}

Cas où le nombre de couples dont les différences non nulles est inférieur ou égal à 20 (n≤20)

Cas où le nombre de couples dont les différences non nulles est inférieur ou égal à 20 ($n \leq 20$)

La distribution de T n'est pas normale; la valeur théorique de T est tabulée

on rejette H0 si T Calculé (T+ ou T-) est inférieur à T lu sur la table.

Dans notre exemple, n = 9 et T calculé = 3,5 (nous avons pris T- parce qu'elle est plus petite que T+); à P = 0,05 T théorique = 6.

L'hypothèse nulle est alors rejetée.

Cas où le nombre de couples dont les différences non nulles est supérieur à 20 (n>20)

La distribution de T tend vers une distribution normale.

$$\text{Sa moyenne est : } m_T = \frac{n(n+1)}{4}$$

$$\text{Son écart-type est : } \delta_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

$$\text{On calcule alors : } |Z| = \frac{|T - m_T|}{\delta_T}$$

La valeur calculée sera comparée à la valeur critique de Z (table de la loi normale réduite).

L'hypothèse nulle est rejetée si la valeur calculée de Z est supérieur à la valeur lue à un seuil donné.

Exemple:

Les données suivantes représentent les notes obtenues par un groupe d'élèves avant et après les vacances. On voulait vérifier l'hypothèse d'une déperdition des acquis:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
Avant	10	12	12	19	5	13	20	8	12	10	8	19	5	11	8	7	4	7	16	2	5
Après	8	10	8	18	8	7	12	10	7	10	3	12	8	11	5	3	5	7	9	8	14

- On commence par calculer D = B-A, ce qui donnera:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
d	2	2	4	1	-3	6	8	-2	5	0	5	7	-3	0	3	4	-1	0	7	-6	-9

On classe les sujets dans l'ordre croissant des valeurs absolues des différences non nulles; on obtient la distribution suivante:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
d	2	2	4	1	-3	6	8	-2	5	0	5	7	-3	0	3	4	-1	0	7	-6	-9

Sujets	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**11**	**12**	**13**	**15**	**16**	**17**	**19**	**20**	**21**
IdI	1	1	2	2	2	3	3	3	4	4	5	5	6	6	7	7	8	9
Rang D	1.5	1.5	4	4	4	7	7	7	9.5	9.5	11.5	11.5	13.5	13.5	15.5	15.5	17	18

$$T^{+} = 120 \qquad T^{-} = 51$$

$$\text{Notons que} : T^{+} + T^{-} = \frac{n(n+1)}{2}$$

$$51 + 120 = \frac{18(18+1)}{2} = 171$$

$$\text{Sa moyenne est : } m_T = \frac{n(n+1)}{4} = \frac{18\times 19}{4} = 85.5$$

$$\text{Son écart - type est : } \delta_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{(18\times 19)(36+1)}{24}} = 22.961$$

$$\text{On calcule alors } |Z| = \frac{|T - m_T|}{\delta_T} = \frac{|51-85.5|}{22.961} = 1.502$$

n compare la valeur calculée de Z à la valeur lue sur la table de la loi normale réduite; Z lue au P=0,05 égale 0,121,

on rejette alors l'hypothèse nulle;

il y a une différence entre les scores des élèves avant les vacances et ceux d'après.

ANALYSE DE LA VARIANCE

L'analyse de la variance (ANOVA) a pour objectif d'étudier l'influence d'un ou plusieurs facteurs sur une variable quantitative.

Nous nous intéresserons ici au cas où les niveaux, ou modalités, des facteurs sont fixés par l'expérimentateur. On parle alors de modèle fixe.

C'est la comparaison de moyennes pour plusieurs groupes (> 2).

Il s'agit de comparer la variance intergroupe (entre les différents groupes) à la variance intragroupe (somme des fluctuations dans

S'il n'y a pas de différence entre les groupes, ces deux variances sont (à peu près) égales. Sinon, la variance intergroupe est nécessairement la plus grande.

L'ANOVA se résume à une comparaison multiple de moyennes de différents échantillons constitués par les différentes modalités des facteurs. Les conditions d'application du test paramétrique de comparaison de moyennes s'appliquent donc à nouveau.

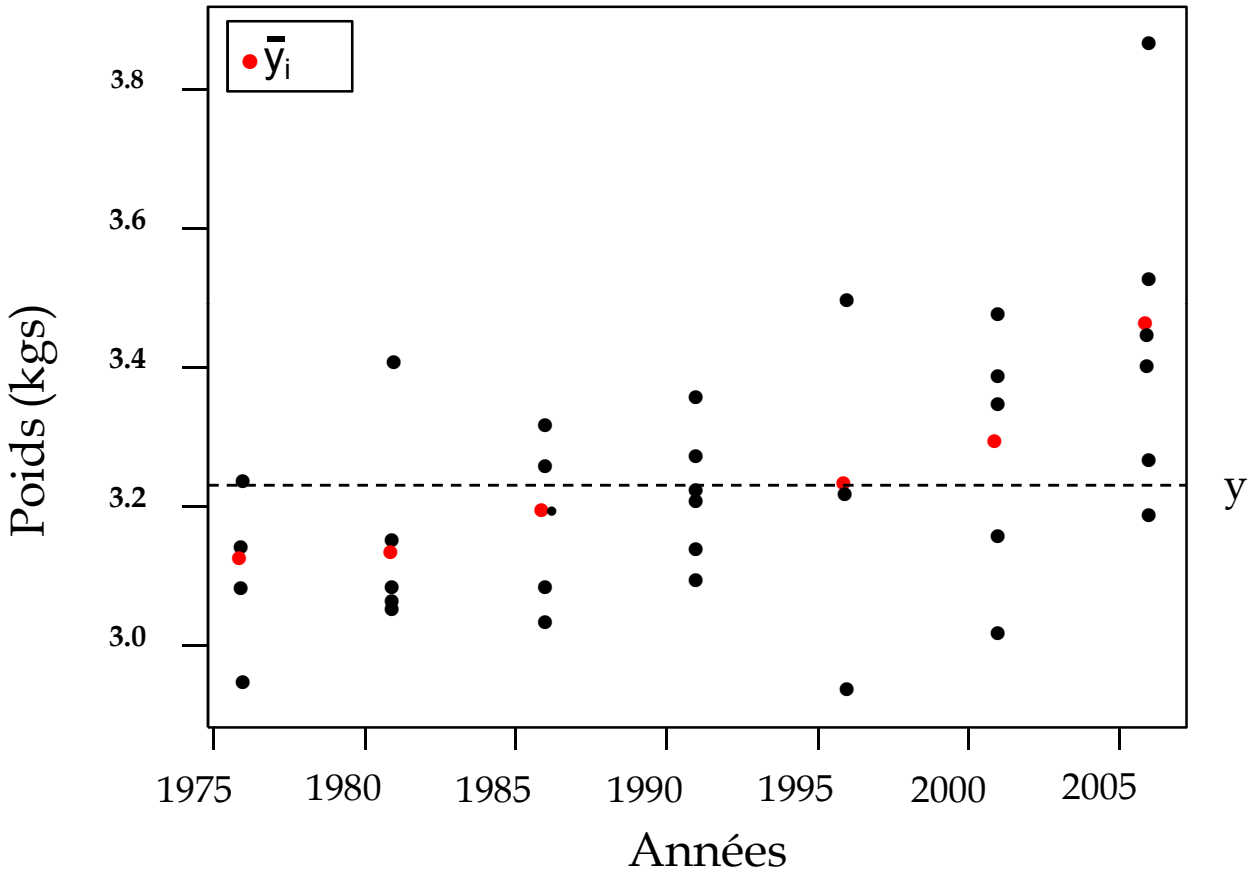
L'analyse de variance (analysis of variance ou ANOVA) peut être vue comme une généralisation du test de Student.

Evolution du poids des marmottes

On dispose des poids de différents individus, au total $N = 37$, pour chaque année.

Année	1976	1981	1986	1991	1996	2001	2006
Poids (kgs)	2.95	2.99	3.07	3.11	2.94	3.34	3.87
	3.24	3.00	3.26	3.26	3.18	3.02	3.41
	3.12	3.41	3.19	3.30	3.50	3.16	3.27
	3.05	3.02	3.32	3.14	3.22	3.48	3.37
		3.05	3.11	3.21		3.39	3.19
		3.13		3.36		3.35	3.53

Anova à un facteur



Anova à un facteur

$$SCE_{totale} = (x_1 - \bar{x})^2 + (x_1 - \bar{x})^2 + ... + (x_i - \bar{x})^2$$

$$SCE_{totale} = (2.95 - 3.23)^2 + (3.24 - 3.23)^2 + ... + (2.99 - 3.23)^2$$

$$SCE_{Inter} = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + ... + n_i(\bar{x}_i - \bar{x})^2$$

$$SCE_{Inter} = 4(3.09 - 3.23)^2 + 6(3.1 - 3.23)^2 + ... + 6(3.44 - 3.23)^2$$

$$SCE_{Intra} = SCE_{totale} - SCE_{Inter}$$

Année	1976	1981	1986	1991	1996	2001	2006
Poids (kgs)	2.95	2.99	3.07	3.11	2.94	3.34	3.87
	3.24	3.00	3.26	3.26	3.18	3.02	3.41
	3.12	3.41	3.19	3.30	3.50	3.16	3.27
	3.05	3.02	3.32	3.14	3.22	3.48	3.37
		3.05	3.11	3.21		3.39	3.19
		3.13		3.36		3.35	3.53
$\bar{x}_i$ (kgs)	3.09	3.1	3.19	3.23	3.21	3.29	3.44
n_i	4	6	5	6	4	6	6

Anova à un facteur

$$ddl_{tot} = N - 1 \qquad ddl_{\text{int } er} = nbr_{col} - 1 \qquad ddl_{\text{int } ra} = ddl_{tot} - ddl_{\text{int } er}$$

$$ddl_{tot} = 37 - 1 = 36 \qquad ddl_{\text{int } er} = 7 - 1 = 6 \qquad ddl_{\text{int } ra} = 36 - 6 = 30$$

$$CM = \frac{SCE}{ddl} \qquad F_{Cal} = \frac{CM_{\text{int } er}}{CM_{\text{int } ra}}$$

Variabilité	ddl	SCE	CM	F_{Cal}	$F_{critique}\,(a = 0.05)$
Totale	36	1.327			
Inter	6	0. 476	0.079		
Intra	30	0.851	0.028	2.821	2.42

On peut donc conclure, au risque de 5%, que le facteur temps a bien un effet sur le poids des marmottes.

Anova à deux facteurs

Les résultats d'une analyse de la variance à deux facteurs avec répétitions sont habituellement présentés dans un tableau comme celui-ci

Analyse de variance à deux facteurs avec répétitions

Source	Somme des carrés	ddl	Moyennes des carres	Fcal
Facteur A	SCFA	I-1	MCFA	MCFA / MCE
Facteur B	SCFB	J-1	MCFB	MCFB / MCE
Interaction	SCAB	(I-1)(J-1)	MCI	MCI / MCE
Résidus	SCR	IJ(K-1)	MCE	
Totale	STC	IJK-1		

Comparaison de trois types d'irrigation

Un agriculteur veut savoir les effets de trois types du système d'irrigation (R1-R2-R3) dans deux type de sols différents (S1-S2).

A partir des échantillons prélèves au nombre de quatre mesure d'humidité dans chaque type de sol (k=1,….4) associe a un type d'irrigation

	Répétition	R1 (j=1)	R2 (j=2)	R3 (j=3)
S1 (i=1)	K=1	43	41	42
	K=2	45	42	44
	K=3	46	43	46
	K=4	53	44	48
S2 (i=2)	K=1	40	35	37
	K=2	40	37	39
	K=3	40	40	40
	K=4	43	40	40

Notre objectif est de tester l'hypothèse d'égalité des moyennes des six échantillons associés à deux facteur (type de sol et type d'irrigation)

$$\bar{x} = \frac{(43 + 45 + ... + 40)}{24} = 42$$

Répétition	$n = 4$
Facteur1	$p = 2$
Facteur2	$q = 3$

	R1 (j=1)		R2 (j=2)		R3 (j=3)		
S1 (i=1)	43 45 46 53	$\overline{x_{i1\,j1}}$ = 46.75	41 42 43 44	$\overline{x_{i1\,j2}}$ = 42.5	42 44 46 48	$\overline{x_{i1\,j3}}$ = 45	$\overline{x_{i1}} = 44.75$
S2 (i=2)	40 40 40 43	$\overline{x_{i2\,j1}}$ = 40.75	35 37 40 40	$\overline{x_{i2\,j2}}$ = 38	37 39 40 40	$\overline{x_{i2\,j3}}$ = 39	$\overline{x_{i2}} = 39.25$
	$\overline{x_{j1}} = 43.75$		$\overline{x_{j2}} = 40.25$		$\overline{x_{j3}} = 42$		$\overline{x} = 42$

$$SCE_{totale} = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(x_{ijk} - \bar{x})^2$$

$$SCE_{totale} = (43-42)^2 + (45-42)^2 + ... + (42-42)^2 = 346$$

$$SCE_A = qn\sum_{i=1}^{p}(\bar{x_i} - \bar{x})^2$$

$$SCE_A = 12 \times [(44.75-42)^2 + (39.25-42)^2] = 181.5$$

$$SCE_B = pn\sum_{i=1}^{p}(\bar{x_j} - \bar{x})^2$$

$$SCE_B = 8 \times [(43.75-42)^2 + (40.25-42)^2 + (42-42)^2] = 49$$

$$SCE_{AB} = n\sum_{i=1}^{p}(\overline{x_{ij}} - \overline{x_i} - \overline{x_j} + \overline{x})^2$$

$$SCE_{AB} = 4 \times [(46.75 - 44.75 - 73.75 + 42)^2 + (42.5 - 44.75 - 40.25 + 42)^2$$

$$+ .....(39 - 39.25 - 42 + 42)^2] = 2.9$$

$$SCE_R = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(x_{ijk} - \overline{x_{ij}})^2$$

$$SCE_R = (43 - 46.75)^2 + (45 - 46.75)^2 + ...(40 - 39)^2] = 112.5$$

$$SCE_T = SCE_A + SCE_B + SCE_{AB} + SCE_R$$

$$pqn - 1 = (p-1) + (q-1) + (p-1)(q-1) + pq(n-1)$$

$$CM_T = \frac{SCE_T}{pqn - 1}$$

$$CM_A = \frac{SCE_A}{p - 1}$$

$$CM_B = \frac{SCE_B}{q - 1}$$

$$CM_{AB} = \frac{SCE_{AB}}{(p - 1)(q - 1)}$$

$$CM_R = \frac{SCE_R}{pq(n - 1)}$$

$$F_A = \frac{CM_A}{CM_R}$$

$$F_B = \frac{CM_B}{CM_R}$$

$$F_{AB} = \frac{CM_{AB}}{CM_R}$$

Corrélation

Concept de la corrélation

La méthode des corrélations a pour but de préciser la relation entre deux séries de phénomènes mesurés, de rechercher la probabilité d'une cause commune ou d'une relation directe de cause à effet entre eux.

Nous le ferons sur un exemple concret, sur le problème forestier suivant:

Y a-t-il relation entre la quantité de pluie tombée en été et l'accroissement du sapin ?

Coefficient de corrélation

Le coefficient de corrélation linéaire r donne une mesure de l'intensité et du sens de la relation linéaire entre deux variables. Son calcul est assez complexe, c'est pourquoi on utilise souvent la calculatrice ou un logiciel.

On s’intéresse ici à son interprétation.

Coefficient de corrélation

Comment interpréter r :

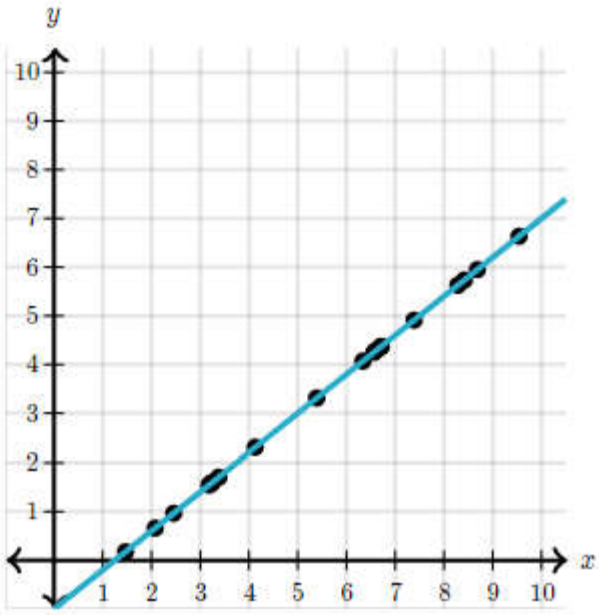
Le coefficient de corrélation est compris entre -1−1minus, 1 et 111.

Plus le coefficient est proche de 111, plus la relation linéaire positive entre les variables est forte.

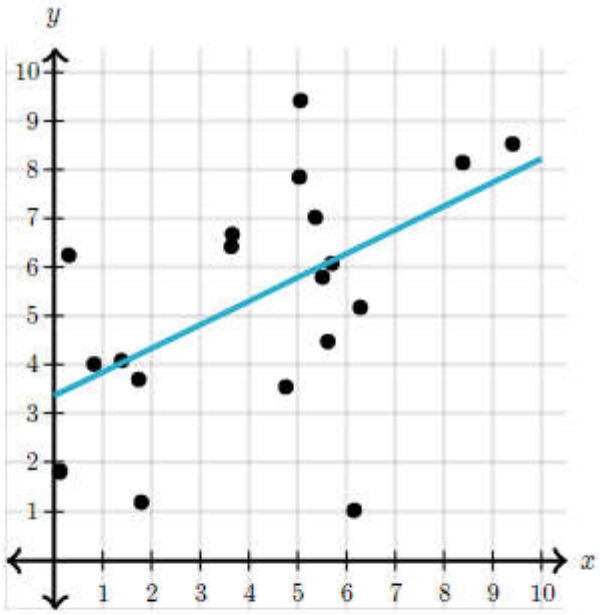
Plus le coefficient est proche de -1−1minus, 1, plus la relation linéaire négative entre les variables est forte.

Plus le coefficient est proche de 000, plus la relation linéaire entre les variables est faible.

Coefficient de corrélation

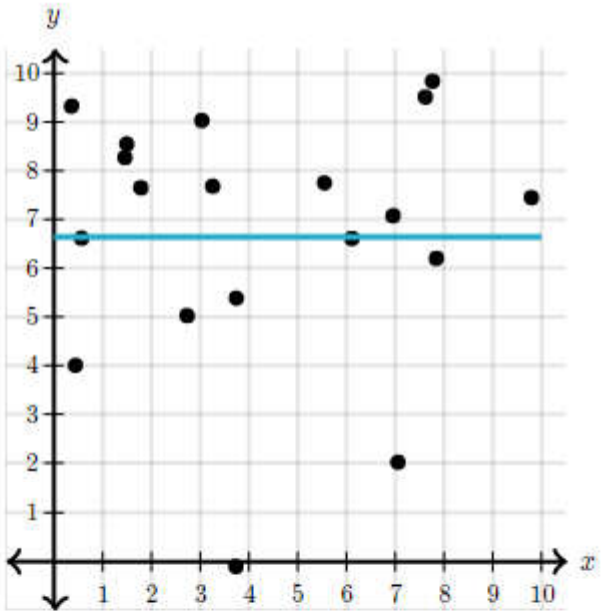


ici, $r = 1$: corrélation positive parfaite entre les deux variables

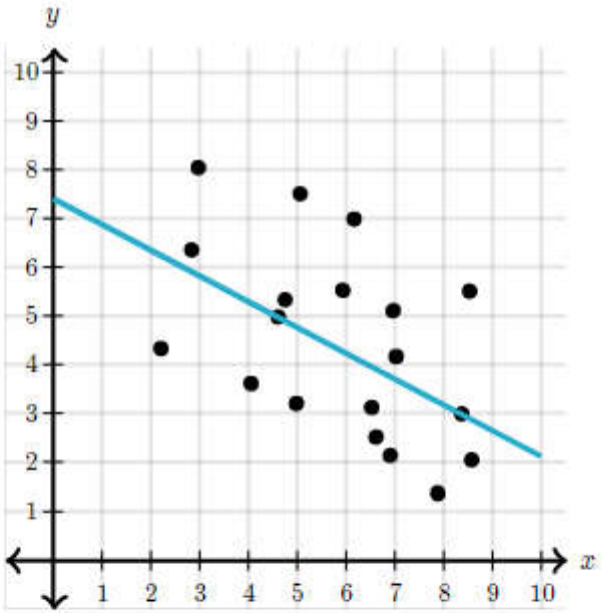


ici, $r = 0.5$: corrélation positive faible entre les deux variables

Coefficient de corrélation



ici, $r = 0$: absence totale de corrélation, les deux variables sont linéairement indépendantes



ici, $r = -0{,}5$: corrélation négative faible entre les deux variables

Coefficient de corrélation

Les notes à l'épreuve de première session d'anglais et de biostatistique de 60 étudiants inscrits en master en 2009 ont été analysées.

Les statistiques descriptives résumées figurent dans le tableau suivant.

Existe-t-il une relation entre la note d'anglais et la note de biostatistique en master ?

	Anglais	Biostatistique
moyenne (m)	13.2	12.7
écart-type (s)	1.5	2.6
Somme (anglais*biostat)	10173	

1. De quel type de problème s'agit-il ?
2. Formulez explicitement les hypothèses du test statistique
3. Quel test statistique utilisez vous ?
4. Quelles sont les conditions de validité de ce test ?
5. Appliquez le test statistique.
6. Que concluez-vous au seuil $\alpha = 0,05$?

1. De quel type de problème s'agit-il ?

- **Corrélation**

- **Tester la liaison entre 2 variables quantitatives :**
  - **note d'anglais**
  - **note de biostatistique**

- **Rôle symétrique**

- **(il est possible que les 2 variables soient liées mais l'une n'est pas susceptible de dépendre de l'autre : il ne s'agit pas d'un problème de régression)**

2. Formulez explicitement les hypothèses du test statistique

Hypothèse nulle (H0) : ρ = 0

Il n'existe pas de liaison linéaire entre la note d'anglais et la note de biostatistique chez les étudiants de master.

Hypothèse alternative (H1) : ρ ≠ 0

Il existe une liaison entre la note d'anglais et la note de biostatistique chez les étudiants de master.

4. Quelles sont les conditions de validité de ce test ?

Liaison linéaire entre les 2 variables Distribution conditionnelle normale et de variance constante

Indépendance des observations

5. Appliquez le test statistique

1. calculez l'estimateur empirique r du coefficient de corrélation

$$r = \frac{Cov(x,y)}{\sqrt{\delta_x^2 + \delta_y^2}}$$

Dont

$$Cov(x,y) = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{n-1}$$

$$Cov(x,y) = \frac{10173 - \frac{(\sum 60 \times 13.2)(\sum 60 \times 12.7)}{60}}{60-1} = 1.9$$

$$r = \frac{1.9}{1.5 \times 2.6} = 0.5$$

5. Appliquez le test statistique

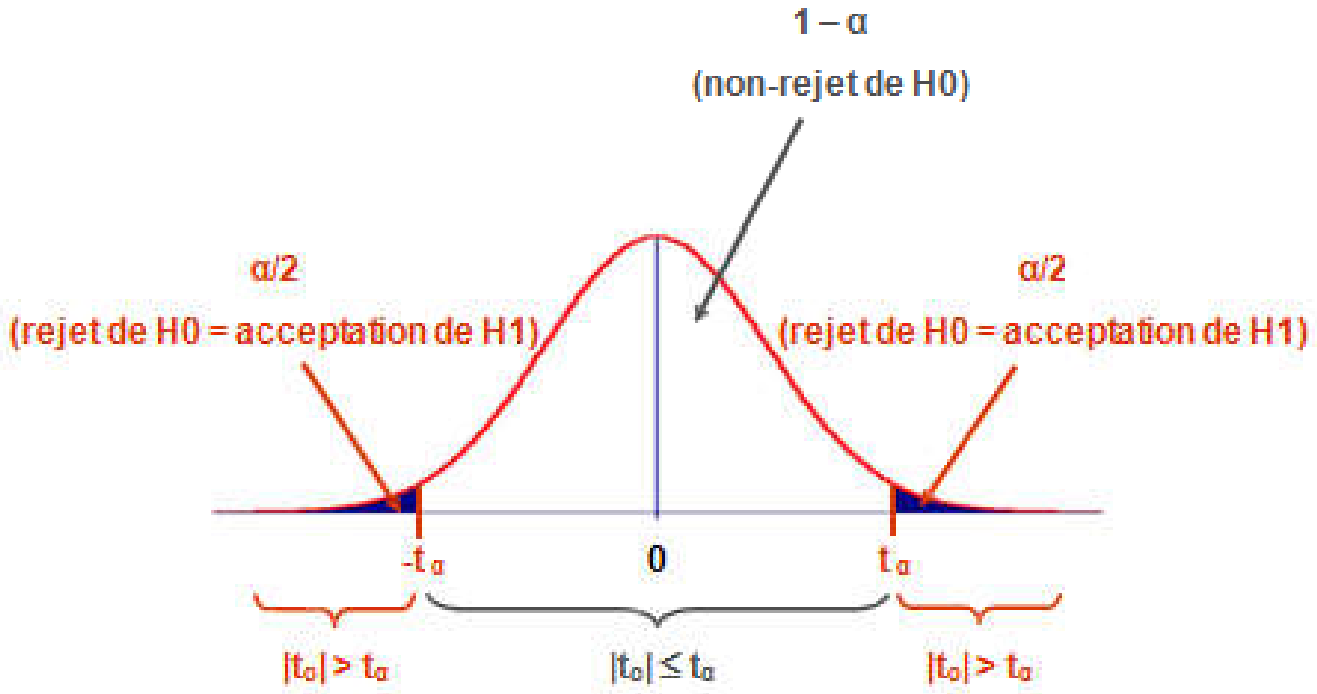
2. calculez la valeur du test du coefficient de corrélation

$$t = \frac{r}{S_r}$$

$$S_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-(0.5)^2}{60-2}} = 0.1$$

$$t_{cal} = \frac{0.5}{0.1} = 5$$

6. Que concluez-vous, avec un risque de 1ère espèce fixé à 0,05 ?



$t_o = 5$ $t_\alpha = 1{,}96$ pour 58 ddl › rejet de H0 : acceptation de H1

Détermination du degré de signification associé à t_o (*P*-value)

- to = 5
- n = 60

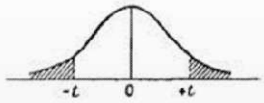
P <0.001

P < α → rejet de H0

Rappel : P-value = probabilité d'observer une valeur de t plus grande que t_o sous l'hypothèse nulle H0

Table de *t* (*).

La table donne la probabilité α pour que *t* égale ou dépasse, en valeur absolue, une valeur donnée, en fonction du nombre de degrés de liberté (d.d.l.).



d.d.l. \ α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,765	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,134	0,741	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,718	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,706	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,703	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,700	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,697	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,695	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,694	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,692	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,691	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,690	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,689	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,688	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,688	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,687	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,686	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,686	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,685	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,685	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,684	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,684	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,684	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,683	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,683	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,683	1,055	1,310	1,697	2,042	2,457	2,750	3,646
∞	0,126	0,674	1,036	1,282	1,645	1,960	2,326	2,576	3,291

(n-2) = 58 ddl →

X

6. Que concluez-vous, avec un risque de 1ère espèce fixé à 0,05 ?

Conclusion

Les notes de 1ère session d'anglais et de biostatistique sont positivement corrélées chez les étudiants de master (r = 0,5, *P*<0,001).

LA REGRESSION LINEAIRE SIMPLE

BUT

TESTER LE TAUX D'INDEPENDANCE ENTRE DEUX VARIABLE QUANTITATIVES

Dans le cas *particulier* où l'on a pu mettre en évidence l'existence d'une *relation linéaire significative* entre deux caractères quantitatifs continus X et Y, on peut chercher à formaliser la relation moyenne qui unit ces deux variables à l'aide d'une des trois équations suivantes :

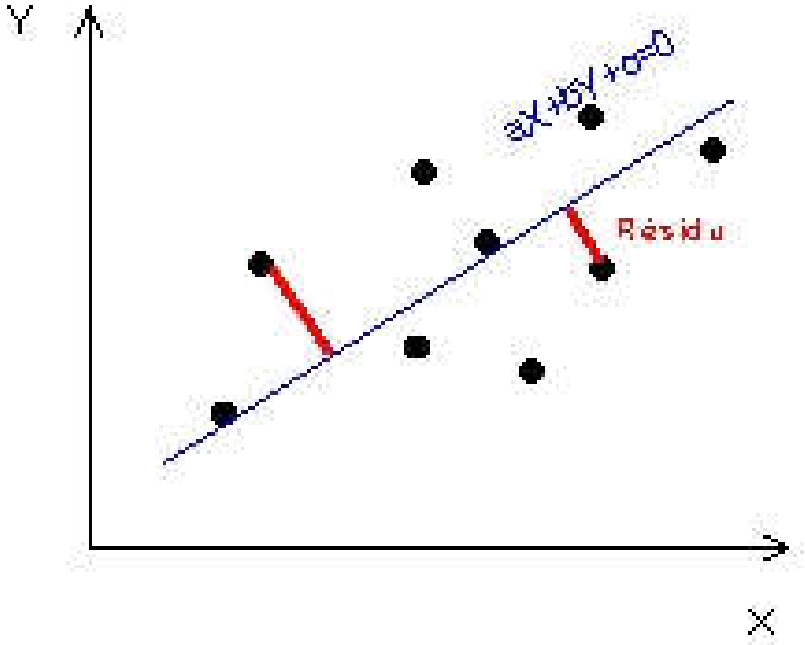
(1) $a.X + b.Y + c = 0$: équation de la **droite moyenne liant les caractères X,Y**

(2) $Y = a.X + b$: **droite de régression de Y en fonction de X**

(3) $X = a.Y + b$: **droite de régression de X en fonction de Y**

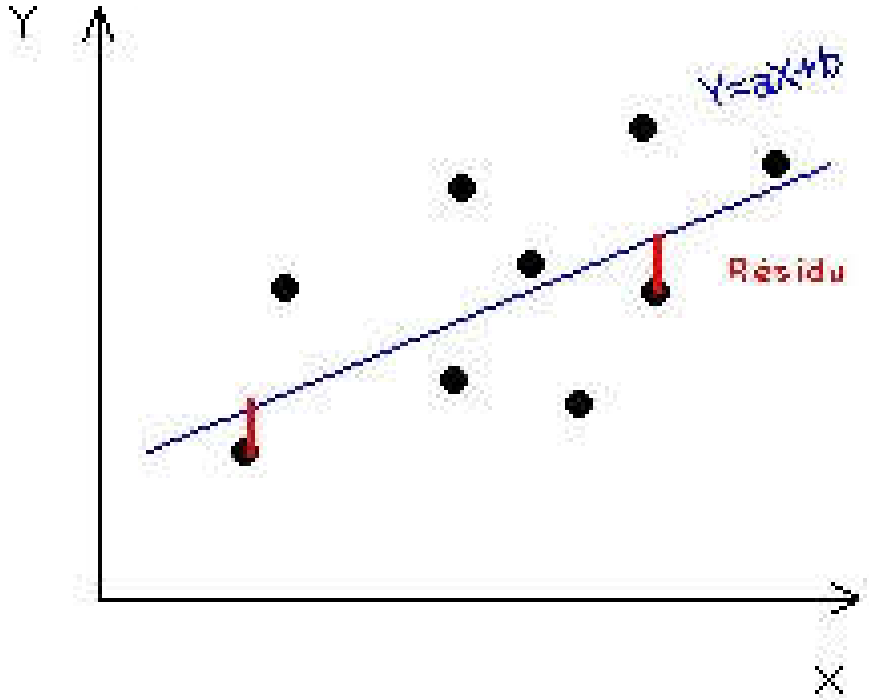
Les trois équations proposées ci-dessus correspondent à trois droites différentes, trois résumés différents du nuage de points (X,Y). La différence entre les trois droites vient du fait que les trois équations proposées correspondent à trois objectifs différents :

(1) **La droite moyenne** est un résumé de la relation entre X et Y qui n'introduit *aucune hypothèse particulière sur le sens de la dépendance causale* qu'il peut y avoir entre les deux variables. Elle visera donc à tracer la droite qui soit *la plus proche de tous les points*, c'est-à-dire les résidus définis par la perpendiculaire de chaque point à la droite moyenne (plus court chemin).



(1) Droite exprimant la relation moyenne entre X et Y

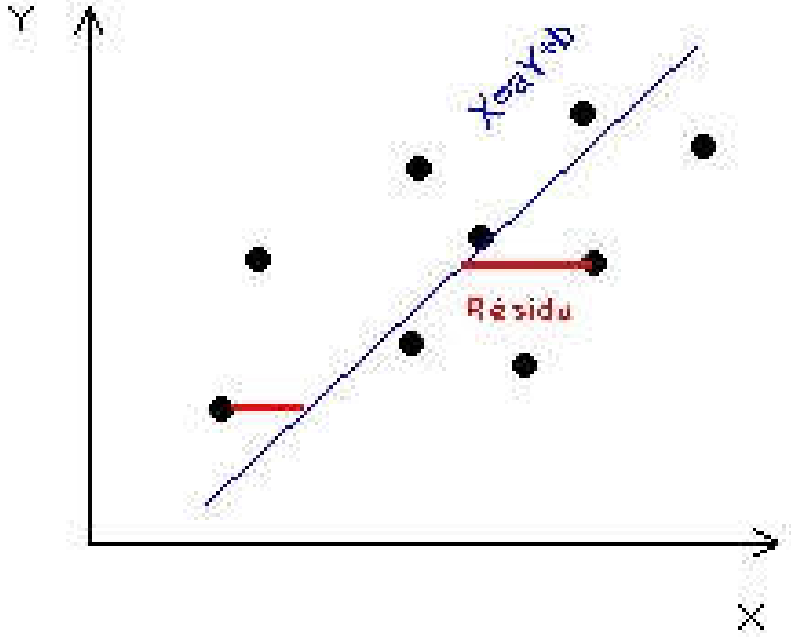
(2) **La droite de régression de Y en fonction de X** introduit l'hypothèse que les valeurs de Y dépendent de celles de X, c'est-à-dire postulent que *la connaissance des valeurs de X permet de prévoir les valeurs de Y*. Il s'agit donc d'un modèle de prévision et l'objectif est de minimiser l'erreur de prévision c'est-à-dire la distance entre les valeurs Yi observées et les valeurs Y*i estimés par la relation Y*=aX+b. Les résidus seront donc la distance à la droite par rapport à l'axe Oy.



(2) Droite exprimant Y en fonction de X

(3) **La droite de régression de X en fonction de Y** introduit l'hypothèse inverse que les valeurs de X dépendent de celles de Y, c'est-à-dire postulent que *la connaissance des valeurs de Y permet de prévoir les valeurs de X*. Il s'agit donc cette fois-ci de minimiser l'erreur de prévision sur X c'est-à-dire la distance entre les valeurs Xi observées et les valeurs X*i estimés par la relation X*=aY+b. Les résidus seront donc la distance à la droite par rapport à l'axe Ox et non plus par rapport à l'axe Oy comme dans le cas précédent.

(3) Droite exprimant X en fonction de Y



7.1 LE CALCUL DE LA DROITE DE REGRESSION Y=aX+b

Un exemple pédagogique de régression linéaire

Pour rendre les choses plus claires, nous partirons d'un exemple simple et très classique qui est celui de la relation entre l'altitude (X) et température (Y) à l'intérieur d'une région de taille suffisamment petite pour que l'on puisse négliger les facteurs de variations macroscopiques de la température (distance à la mer, latitude, etc.).

Les données présentées sur la Figure 2 sont imaginaires mais elles pourraient correspondre à la situation d'une vallée alpine de direction nord-sud pour laquelle on a procédé au relevé des températures à midi dans huits stations situées à des altitudes différentes et localisées sur chacun des versants de la vallée.

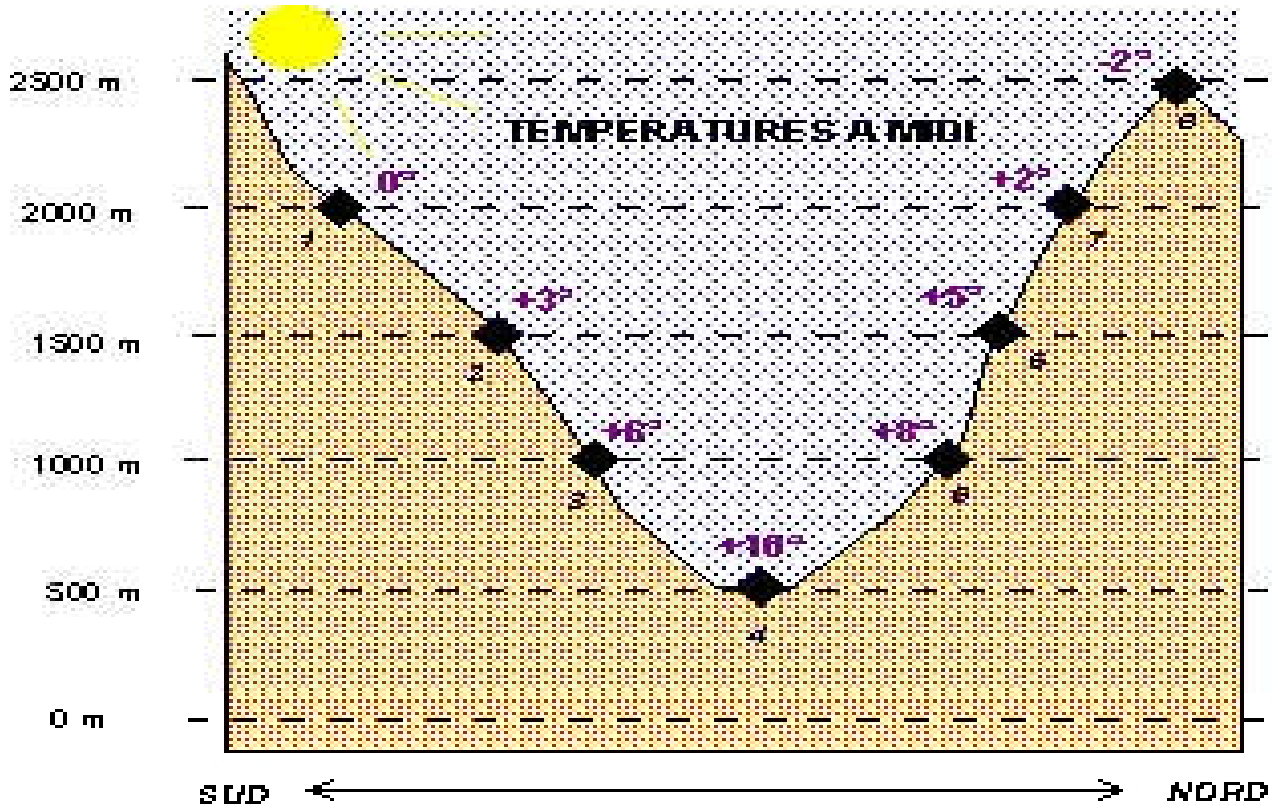


Figure 2 : Température et altitude dans 8 stations (données imaginaires)

Les données relatives à l'altitude et la température des 8 stations peuvent être rassemblées dans un tableau (tableau 1) à partir duquel on calculera les paramètres caractériques de chaque variable (moyenne et écart-type) ainsi que leur covariance.

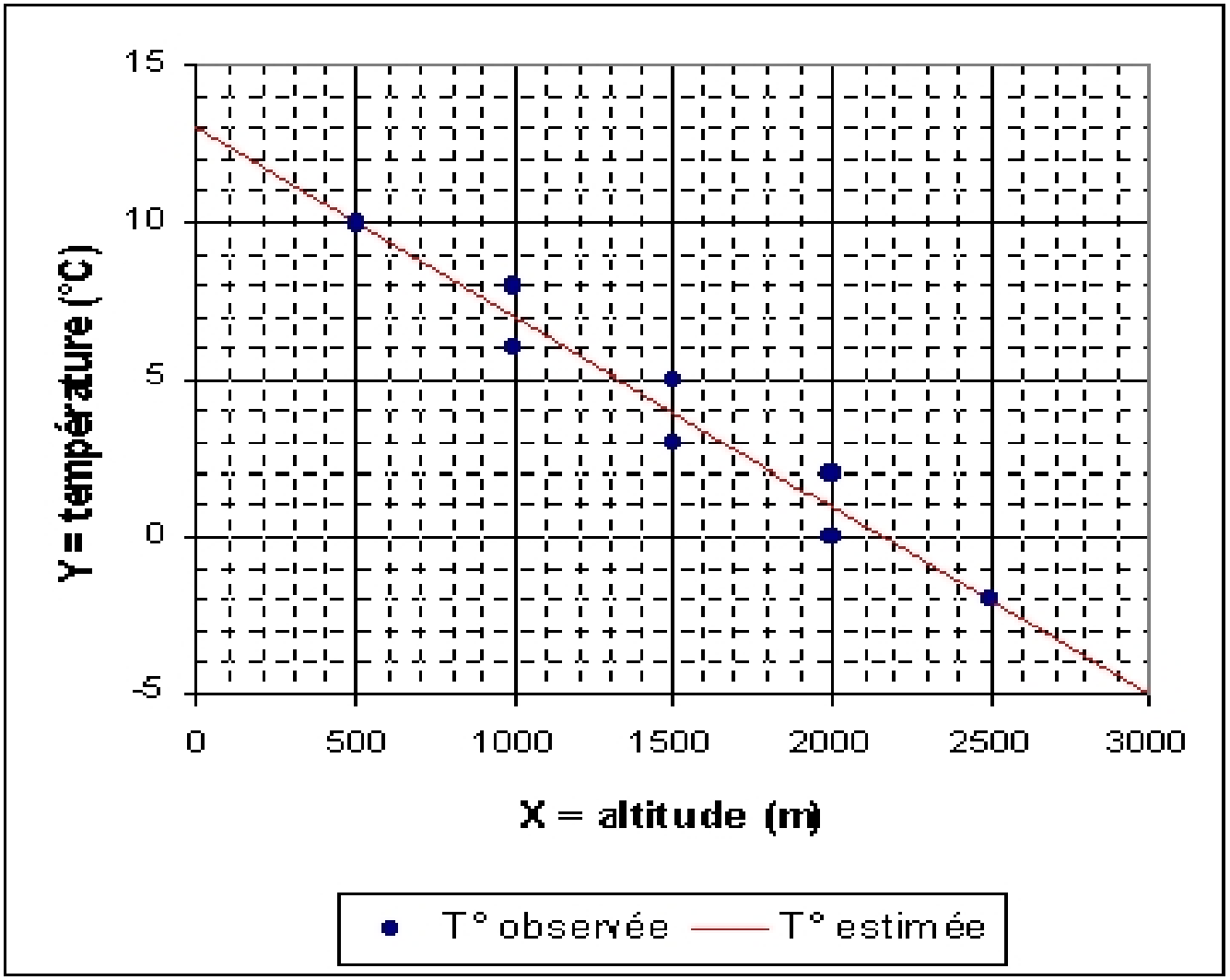
Tableau 1 : Paramètres caractéristiques de la température (Y) et de l'altitude (X) de 8 stations météorologiques (données imaginaires)

i	(Xi) Altitude (m)	(Yi) Température (°)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	2000	0	500	-4	-2000
2	1500	3	0	-1	0
3	1000	6	-500	2	-1000
4	500	10	-1000	6	-6000
5	1000	8	-500	4	-2000
6	1500	5	0	1	0
7	2000	2	500	-2	-1000
8	2500	-2	1000	-6	-6000
moyenne	1500	4	0	0	-2250
écart-type	612	3.8	-	-	-

On déduit de la valeur de la covariance (-2250) et de celle des deux écarts-type (612 pour X et 3.8 pour Y) l'existence d'une très forte corrélation linéaire négative entre les deux variables :

r(X,Y) = Cov(X,Y) / [ect(X) * ect(Y)] = -2250 / (612 * 3.8) = -0.97.

Détermination de la droite de régression par le critère des moindres carrés



Nous avons vu en introduction que lorsque l'on cherche à exprimer Y en fonction de X, on peut affecter à chaque valeur observée Y_i une valeur estimée par la droite de régression $Y^*_i = aX_i+b$. L'erreur d'estimation sur l'individu i est donc égal au résidu ε_i défini par :

$$\varepsilon_i = (y_i - y_i^*) = y_i - (ax_i + b)$$

Comme on souhaite obtenir un ajustement global qui soit optimal pour l'ensemble des stations, il faut définir un critère général définissant la qualité d'ajustement de l'ensemble des valeurs à la droite proposée.

(a) La première solution (ERR1) qui vient à l'esprit est la **minimisation de la somme des résidus** :

$$Err1 = \sum \varepsilon_i$$

Mais ce critère est à l'évidence discutable car des résidus positifs et négatifs peuvent se compenser (températures sous-estimées ou sur-estimées par le modèle) et l'on pourrait obtenir un ajustement optimal (ERR1=0) alors même que la droite ne passerait pas par tous les points du nuage.

(b) La seconde solution (ERR2) consiste alors évidemment la **minimisation de la somme des valeurs absolues des résidus :**

$$Err2 = \sum \left| \varepsilon_i \right|$$

Il s'agit cette fois-ci d'un bon critère, mais qui a le défaut de ne pas posséder de solution analytique et qui impose une recherche itérative sur toutes les droites du plan.

(c) La troisième solution (ERR3) qui est la plus souvent retenue en statistique est appelée **critère des moindres carrés** et consiste à **minimiser de la somme des carrés des résidus :**

$$Err3 = \sum (\varepsilon_i)^2$$

Comme dans le cas précédent, le critère est correct puisqu'il n'y a pas de compensation entre les résidus positifs et négatifs et la valeur ERR3 ne s'annule que si tous les points du nuage sont alignés le long d'une droite. Mais ce critère possède l'immense avantage d'aboutir à une solution analytique très simple. En effet l'équation de la droite de régression Y=aX+b qui minimise le carré des écarts entre valeurs observées et valeurs estimées est obtenue très simplement à l'aide des deux formules suivantes :

Les valeurs optimales d'ajustement des paramètres de la droite Y=aX+b ***pour le critère des moindres carrés*** *sont données par les relations :*

$$a = Cov(x, y) \Big/ \delta_x^2$$

$$b = \overline{y} - a\overline{x}$$

Appliquées aux données du Tableau 1, ces équations permettent d'obtenir les paramètres optimaux d'ajustement de la droite de régression de la température en fonction de l'altitude :

a = -2250 / (612*612) = -0.006 (°C / m)

b = 4 - (-0.006 * 1500) = 13 (°C)

Température (°C) = -0.006 * altitude (m) + 13

$Var(y)$	$=$	$Var(y^* = ax+b)$	$+$	$Var(\varepsilon)$	
information totale	$=$	information modélisée	$+$	information résiduelle	

La qualité de l'ajustement correspond donc au rapport entre l'information totale sur Y et l'information effectivement reconstituée à partir de la connaissance procurée par la variable X. Cette qualité d'ajustement varie entre 0% (X n'apporte aucun élément de prévision sur Y) et 100% (la connaissance des valeurs de X permet de prévoir intégralement les valeurs de Y) et dépend de l'intensité de la corrélation entre X et Y. Elle peut se calculer (Tableau 2) ou se mesurer directement à l'aide du coefficient de détermination, c'est-à-dire du carré du coefficient de corrélation entre X et Y. Si l'on opte pour le calcul, on constate que la variance des températures observées (16.2) est bien égale à la somme de la variance des températures estimées (15.4) et de la variance des résidus (0.9). La qualité d'ajustement est donc égal à 15.4/16.2 soit 0.95 ce qui correspond également au carré du coefficient de corrélation linéaire des variables X et Y : $(-0.97)^2 = 0.95$.

$$\text{Qualité d'ajustement} = \frac{\text{Var}(y^*)}{Var(y)} = [r(x,y)^2] = coef.déter$$

Analyse des résidus d'une régression linéaire

Même si l'importance des résidus d'un modèle de régression est limitée, il est toujours instructif de procéder à leur analyse afin de vérifier :(1) si les résidus ne révèlent pas une mauvaise spécification du modèle utilisé
(2) si les résidus ne mettent pas en évidence l'existence d'autres variables explicatives que celle qui a été retenue.

Le premier point sera développé ultérieurement (régression non-linéaire) et l'on va se contenter de développer ici le second à l'aide de l'exemple des températures et de l'altitude.

Tableau 2 : Analyse des résidus de la régression : Température = -0.006 Alt + 13

i	(Xi)	(Yi)	Y*i=aXi+b	Yi-Y*i
1	2000	0	1	-1
2	1500	3	4	-1
3	1000	6	7	-1
4	500	10	10	0
5	1000	8	7	1
6	1500	5	4	1
7	2000	2	1	1
8	2500	-2	-2	0
moyenne	1500	4	4	0
variance	428571	16.3	15.4	0.9

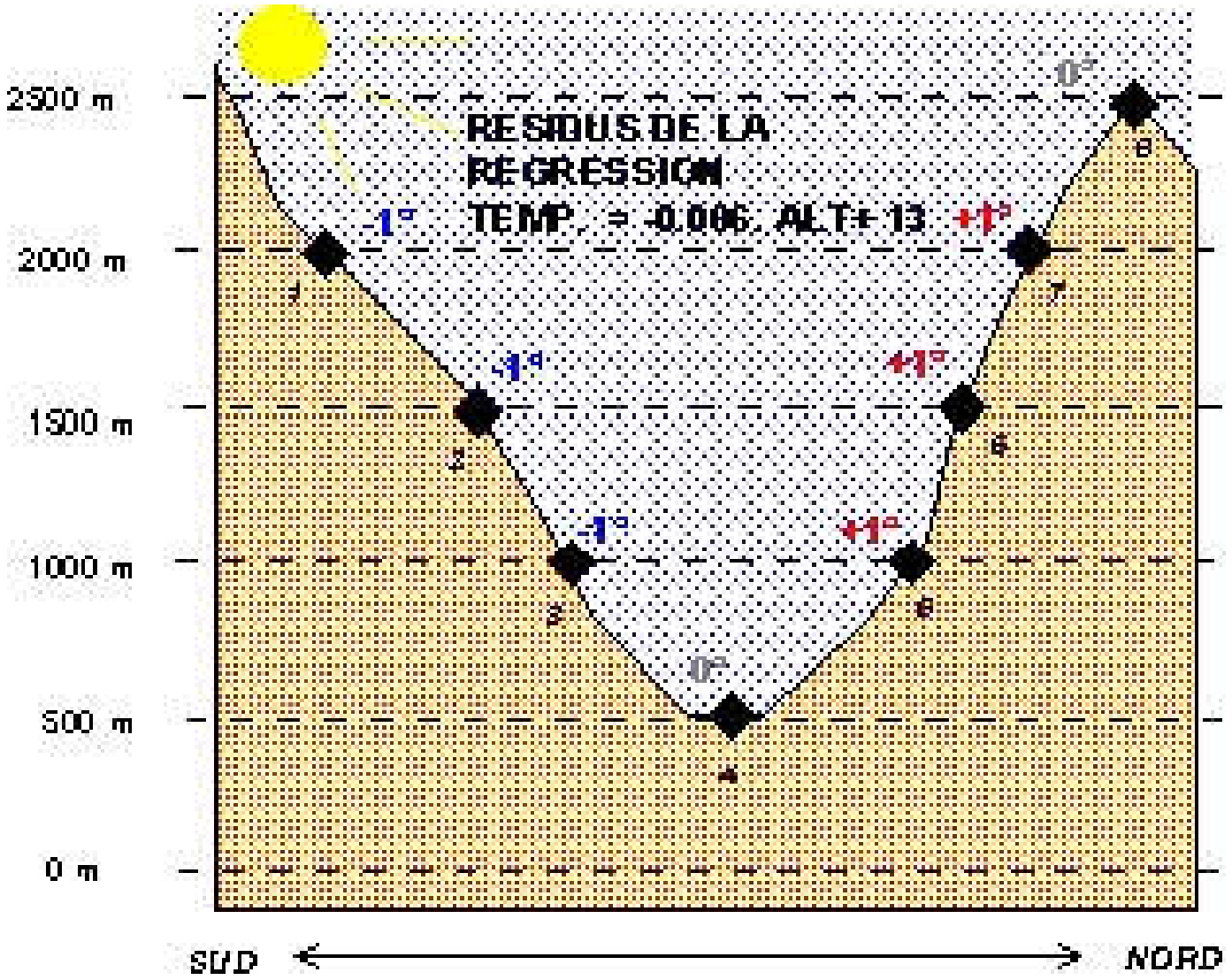
La somme des résidus est nulle (**propriété de la régression linéaire**) mais on constate que trois stations ont des résidus positifs (température réelle supérieure de 1° à ce que laisserait prévoir leur altitude) et trois autres des résidus négatifs (température réelle inférieure de 1° à ce que laisserait prévoir leur altitude).

Comment interpréter ces écarts ?

On peut tout d'abord supposer que ces écarts sont une composante aléatoire liée à l'imprécision des outils de mesure utilisé (précision des thermomètres) mais dans ce cas là la disposition spatiale des résidus devrait être aléatoire. Or, lorsque l'on cartographie les résidus (Figure 4) on constate que les résidus positifs et négatifs sont loin de se disposer au hasard dans l'espace.

Figure 4 : Configuration spatiale des résidus de la régression :

Température = -0.006 Alt. +13



APPLICATIONS PRATIQUES DE LA REGRESSION LINEAIRE

Problème : supposons que l'on dispose de 100 stations météorologiques en Auvergne, pour lesquelles on mesure l'altitude en mètres (X) et la température moyenne (Y) tout au long de l'année.

Est-il réellement utile de retenir chaque jour les 100 températures ?

Réponse : l'observation a montré qu'il y avait une forte corrélation négative (-0.97) entre l'altitude et la température.

La droite de régression T°C = -0.006.Am+10° permet de résumer l'essentiel de l'information sur la variation spatiale des températures (-0.9*-0.9= 81%), dès lors que l'on connaît l'altitude.

Conclusion : la régression permet de résumer un ensemble volumineux d'informations à l'aide de deux paramètres. Ce résumé est évidemment d'autant plus valable que la corrélation est élevée.

Modéliser

Problème : l'observation répétée tout au long de l'année montre que le coefficient a ne change guère (-0.006) alors que le coefficient b varie selon les saisons (élevé en été et faible en hiver) : que peut-on en déduire ?

Réponse : le coefficient a indique de combien varie la température chaque fois que varie l'altitude. Ainsi, une variation d'altitude de +100 m correspond à une diminution de température de -0.006*100 = -0.6°C : c'est le **gradient thermique.** Le coefficient b indique la t° correspondant au cas ou l'altitude est de 0m : c'est donc la **température moyenne ramenée au niveau de la mer**.